FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Predicting drug effectiveness in Cancer Cell Lines using Machine Learning and Graph Mining

Diogo Nunes

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Rui Camacho (FEUP)

Predicting drug effectiveness in Cancer Cell Lines using Machine Learning and Graph Mining

Diogo Nunes

Mestrado Integrado em Engenharia Informática e Computação

Resumo

O cancro é uma doença heterogénea, com um nivel de diversidade entre tumores considerável. Os biomarcadores, no contexto de uma doença oncológica, permitem a identificação da capacidade de resposta de um paciente a um dado fármaco. Estes tratamentos especificos têm produzido resultados em média superiores aos de uso mais abrangente. No entanto a ligação entre a resposta ao tratamento e o valor de um dado biomarcador é em muitos casos ainda desconhecida. O objectivo deste projecto é, com base em resultados prévios e na caracterização tanto dos fármacos como dos tecidos celulares, conseguir prever a eficácia de um fármaco em um tumor utilizando técnicas de Graph Mining e Machine Learning. Um modelo anterior será usado como base, começamos por replicá-lo e experimentamos com a sua transformação num problema de classificação e com a introdução de substructuras moleculares dos fármacos encontradas através do uso de Graph Mining. Os resultados indicam uma performance aceitável.

Abstract

Cancer is an heterogeneous disease, with a high degree of diversity between tumours. Biomarkers, in the context of an oncological disease, allow the identification of the response from a patient to a given drug. These specific treatments have been producing results that are superior on average to broader ones. However, the relationship between a drug's response a biomarkers value is in many cases yet unknown. Some models to predict this relationship have already been built, using machine learning methods. The input are characterizations of both the drug and the tissue along with the result of the drug's use on a given tissue. The goal of this thesis is to improve on previous models and the characterization of both the drug and the tissue through the introduction of graph mining and other machine learning methods. A previous model will be used as a baseline, we start by replicating it and we experiment with it's transformation into a classification problem and through the addition of drug molecular substructures found using Graph Mining. The results indicate acceptable performance.

Agradecimentos

I would like to thanks my family, friends and my supervisor for their companionship during this project.

Diogo Nunes

"You should be glad that bridge fell down.
I was planning to build thirteen more to that same design"

Isambard Kingdom Brunel

Contents

List of Figures

List of Tables

Abbreviations and Symbols

ANN	Artificial Neural Network
FSM	Frequent Subgraph Mining
GM	Graph Mining
IC50	Inhibitory Concentration 50
ML	Machine Learning
MSE	Mean Square Error
NCI60	National Cancer Institute 60
PaDEL	Pharmaceutical Data Exploration Laboratory
R	Pearson Correlation
RF	Random Forests
RMSE	Root Mean Square Error
SMILES	Simplified Molecular-Input Line-Entry System

Chapter 1

Introduction

This project, in it's simplest description, consists of trying to predict the effectiveness of a drug on a tissue using machine learning. By using mathematical representations of both cell lines, drugs as inputs it is possible to predict the inhibitory values. Other projects have followed this approach, we will start by replicating their steps, analyze them and add and analyze graph mining concepts. In the following sections the motivation, goals and document structure will be described.

1.1 Motivation

Cancer is currently one of the leading causes of death in the world with an annual number of cases of 14 million and a mortality rate of about 50% [1]. It is expected that the number of cases rise 70% over the next two decades. The effects of both the disease and the treatments are also associated with a deterioration of the quality of life of the affected individual [2]. No one size fits all cure for it has been found so far and treatment relies mostly on chemotherapy.

Cancer has long been known to be a heterogeneous disease with a diverse number of tumour profiles - each tumour is made of cells with a different genetic profile. Therapies that can target specific profiles, named targeted therapies, have shown superior results contributing to the growing belief that targeting specific profiles will allow for more efficient treatments [3]. Examples are Trastuzumab, which targets the *ERBB2* biomarker and has been shown to increase the survivor rate of early breast cancer treatments by 3.3% [4], and Imatinib targeting *BCR-ABL* [3]. However, approved drugs remain at a low number and their usage in treatments remains scarce. The relationship between biomarker values and a drug effectiveness remains mostly unknown and it is cumbersome and costly to experiment on all possible tissues which is why a computerized model that could predict this relationship would be valuable.

Cultures of cancer cells have been developed by laboratories through the growth of human-derived tumour cells *in vitro*. These cell lines can then be used to reliably test the effects of drugs as they have the same genetic characteristics as the original tissue. This has enabled high throughput testing resulting in a large amount of collected data.

The use of Machine Learning (ML) and Data Mining enables the processing of larger amounts

of data for pattern finding without explicitly programming them. It has been used in this particular field, it's use in the prediction of drug efficiency in cancer cells has been 20% more accurate than statistical methods [5]. Using available previously performed cell line experiments, with cell line data and drug molecular information it's possible to create a model to predict future drug reactions. There are a number of existing models following this approach [5]–[7] each using a different dataset. Most achieved encouraging results and reasonable accuracy but still falling short of providing a replacement for laboratory experimentation.

1.2 Goals

The goal of this project is to better understand the relationships between cancer cells and the effectiveness of a drug in order to improve the drug's administration and provide more specific treatments. This goal will be achieved through the construction of a model that can predict the effectiveness of a drug. Besides being able to predict the effectiveness it's also important that we can acquire information about the inputs, results and effects achieved. We will first replicate Menden *et al.* experiments which uses cell line and IC50 values from the Genomics in Drug Sensitivity project with added drug information. After analyzing the model, sub-molecules found through Frequent Subgraph Mining will be added and their performance also analyzed.

1.3 Document Structure

Beyond this introductory Chapter 1, this dissertation's Chapters are structured in the following way. The second Chapter 2 details the relevant state of the art topics. In the next Chapter 3 the methodology for the experiments performed is detailed with the results discussed in Chapter 4. Finally, in the final Chapter 5, the overall goal satisfaction and future work possibilities are discussed.

Chapter 2

Biology and Data Mining Concepts

This chapter provides an overview of the state of the art focusing on the relevant subjects for this dissertation. The first section contains brief explanations of biology concepts necessary for the understanding of the work done, followed by a similar overview for cheminformatics. On the third section a description of the Data Mining process and concepts can be found. The chapter concludes with a description of relevant machine learning concepts and algorithms.

2.1 Biology Concepts

Cancer

Cancer is a group of diseases originated from the uncontrolled and abnormal growth and multiplication of cells [8]. The process comes from the failure of a cell to replace itself and instead divide indefinitely. These cells will then, except in blood cancer (leukemia) form a tumour. This phenomenon can happen on any of the body's organs, the cancer's name usually takes after organ's (ex. Brain Cancer). If left unchecked the cell growth will impair the functioning of local organs, cause body-wide issues such as fatigue and fever and may eventually lead to death.
Cancer is a major disease affecting a large number of people. It is is the second leading cause of death worldwide, in 2012 14 million contracted cancer and in 2015 it caused 8.8 million deaths [9]. The aberrant cells that compose the tumour go through a unique combination of genetic changes making them different for every patient. In addition the cells and blood vessels around the tumour, known as it's microenvironment, influence the tumour and vice-versa. These innumerous possible changes and environments mean that cancers are heterogeneous diseases [10] [11].

Biomarkers

There is no clear definition of what a biomarker is. The broadest definition is by the Center for Biomarkers in Imaging which considers a biomarker any "physiological, biochemical or molecular parameter associated with the presence and severity of specific disease states" [12]. Blood pressure or urine samples are examples of what can be considered a biomarker under this definition. They

can perform one of three functions - *diagnostic biomarkers* can help diagnose a condition, *prognostic biomarkers* help predict the diseases' aggressiveness and *predictive biomarkers* are used to forecast a treatment's effectiveness. Predictive and prognostic biomarkers are key for identifying drugs that can target a tumours' specific characteristics and achieve higher therapeutic performance.
In cancer they are used to characterize the tumour and its microenvironment attempting to identify it's driving mutations [3], allowing for more targeted therapies. They can be found at either the molecular, cellular or tissue level.

Cell Culture

Cell cultures are cells grown in artificial environments (*in vitro*) by supplying them in "the appropriate nutrients and conditions" [13]. They are used to further study the cell and test drugs, among others.

Cell lines

A cell from a culture has a limited life span. To prevent their death and to allow easier experimentation they can be subcultured (passaged) into a cell line [14]. The main advantage is their lower maintenance and cost of use, the negatives include the dangers of contamination between different cultures in a single line and a less accurate representation of the original tissue. There are some doubts about their ability to evaluate a drug's effectiveness but there is a scientific consensus that they presently are the best possible tool to evaluate it [11]. Cell lines have a long use history in cancer research, the National Cancer Institute 60 (NCI60) introduced cell line screening to cancer drug testing. It was developed in the 1980s and contained 60 cancer cells lines for testing. It is considered a scientific landmark in cancer research, a number of new technologies were developed and applied that remain cornerstones of cell line screening. Interest in Cancer cell line research has been reinvigorated due to the increased understanding of cancer's heterogeneity. Which also made the need to increase the number of cell lines it clear so more diverse sets of cells can be represented. Today there are cancer cell lines available with a higher number of cell lines that can be used to test targeted therapeutics.

IC50

The inhibitory concentration 50 (IC50) is a common measure of a drug's effectiveness. It represents the amount of an inhibitor that is required to halve the effectiveness of a biological process [15]. It's ubiquitous use is due to the lower variance from experience environments and easily calculable standard error [16].

Genome

The genome is the genetic material contained inside an organism [17]. It is mostly comprised by the Deoxyribonucleic acid (DNA), which contains the instruction set for an organism to develop. Every cell contains a copy of the DNA and analysis has shown it's influence in cancer tumour formation [18]. Some DNA concepts that are relevant to the project are introduced in the following paragraphs.

Copy Number

The copy number is the number of repetitions in certain DNA sections. Around 5 to 10% of human DNA is comprised of repetitions [19], these repetitions or deletions can help characterize a genome especially since copy number anomalies are more common in tumour cells due to their frequent mutations [18] .

Phenotype

The phenotype are the "observable characteristics" of an organism due to it's gene expression, according to the National Cancer Institute [20]. The wildtype phenotype refers to the phenotype of an organism as seen in nature, in other words - it's standard version. Other variants are referred to as mutations. In cancer, since cells undergo a significant number of mutations, analyzing the mutator phenotype can provide insight into treatment possibilities [21].

Microsatellite Instability

Microsatellite instability (MSI) is a phenomenon caused by a damaged mismatch repair (MMR) system in the DNA. It's presence has proven to be a reliable cancer predictor [22]. The DNA MMR is a system that fixes DNA errors during replication and recombination, if it is faulty then cells are likely to accumulate errors which makes them more prone to abnormal behavior.

2.2 Cheminformatics

Cheminformatics is the application of computer technology to chemistry problems [23]. It entails the extraction of information from chemical process through computer - *in-silico* - models.

Quantitative structure–activity relationship

Quantitative Structure–activity relationship (QSAR) is a name given to computer models that attempt to predict the link between a molecular structure and it's biological activity [24]. Its process encompasses the representation of a molecule in a digital readable representation, and model construction, usually using data mining and statistical analysis methods, and model evaluation [23].

Molecular Descriptors

Molecular descriptors [25] are mathematical representations of the structure of a molecule in a way that allow statistical computation or for other information technologies to be used on the characterization of the molecule. They can be either obtained by experimental measurements or can be derived through symbolic representation of the molecule. Molecular descriptors can be categorized by the dimensions they describe. 0D-Descriptors merely contain information derived from the chemical formula of the molecule such as the number of atoms, 1D-Descriptors are based on the structural representation and fragments of that structure, mass and bond types for example. 2D-Descriptors are based on a graph model which allows a topological representation of the molecule from which we can derive measures such as graph connectivity, as an example. 3D-Descriptors adds the spatial representation of the molecule creating a full geometric description. 4D-Descriptors also include molecular interaction fields and electron distribution.

Chemical File Formats

There are several file types for molecular representation, the most relevant for this thesis are introduced in the following sections.

SMILES

Simplified molecular-input line-entry system (SMILES) is a specification for representing molecules as ASCII strings. For example Di-nitrogen, that has two nitrogen elements connected through a triple bond, would be represented as N#N.

Chemical table files

Chemical table files are formats for representing the information contained in tables. SDF is a format for representing chemical information. SDF stands for structured data format, it contains the structure of the molecule in addition to some general information about it[26].

Available Tools

There are free available tools that can be used to compute molecular descriptors. Open Babel [27] is a software package that can calculate a molecule's fingerprint and also convert chemical files between formats. It provides bindings for several formats and is licensed under the GNU General Public License. To calculate molecular descriptors PADEL-descriptor [28] is a popular choice. This software can handle the computation of most molecular descriptors, through the use of its own algorithms and by wrapping the Chemistry Development Kit [29].

2.3 Data Mining

Data Mining [30] is the process of extracting knowledge from unorganized and large amounts of data. The basic steps of the process involve the cleaning, integration, selection and transformation of the data; these first steps are typically named the preprocessing steps. The main step is the mining itself where methods, usually involving machine learning, are applied to the data to extract the information in the data, typically patterns. Finally, the results are evaluated to determine the usefulness of the patterns found and presented. A simple version of the process is illustrated in figure 2.1. It's frequent for these steps to then be combined in more complex combinations. The

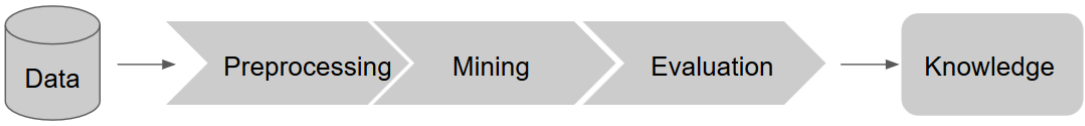

Figure 2.1: Overview of a basic Data Mining process

Cross Industry Standard Process for Data Mining (CRISP-DM) [31] is a common process model used in Data Mining. It formalizes the additional step of understand the data and it's context before preprocessing while also considering the process an iterative looped process where the model is reanalyzed after the evaluation is performed. Figure 2.2, taken from IBM's knowledge center [32], represents this process.

Data Preprocessing

Data preprocessing [30] encompasses the process of treating data so that it fits to a certain standard of quality before it is used by an application. Big data warehouses with data from various sources are prone to bad quality resulting from missing or inaccurate values or different formats among others. The negative factors affecting the data we are focused on removing in this preliminary step are the lack of accuracy, completeness, consistency and interpretability.

Preprocessing can be divided in several steps: Data Cleaning where routines are created to fill missing values and smooth the data. Data reduction where the original data set is either compressed or split while trying to maintain the insights of the same data set, and data integration which consists maintaining the quality of the data while introducing a new data set.

Data cleaning deals with the inconsistency, noise and missing values usually by the creation of routines that run regularly. To deal with missing values we have several options. The simpler ones are ignoring the data record or filling the value manually. There are more complex ways of filling the data, we can use the median of the values of other tuples, additionally we can run a classification algorithm beforehand and use only the objects of the same class to calculate the median. It is also possible to use machine learning algorithms to calculate the most probable value.

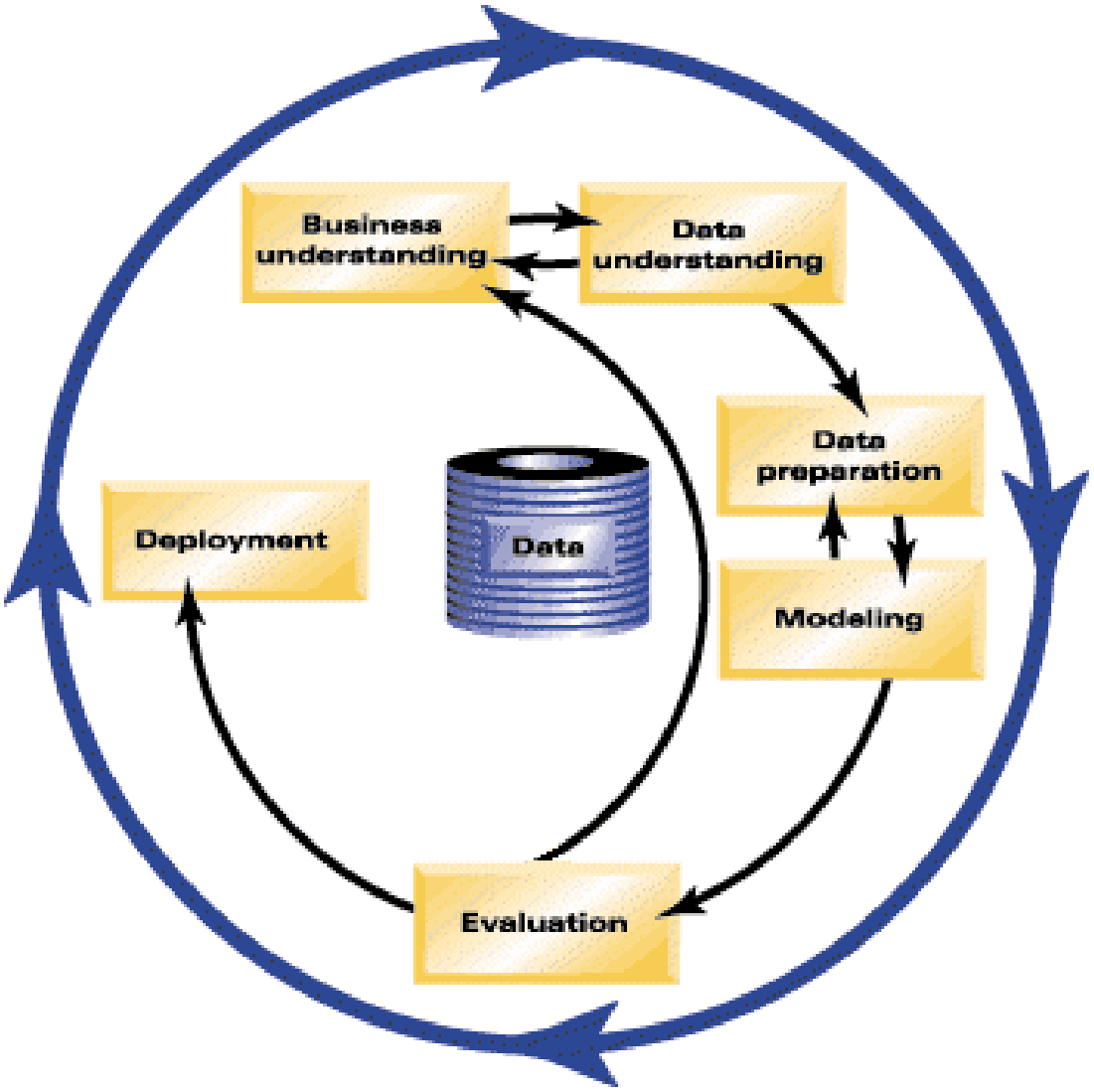

Figure 2.2: Overview of the CRISP-DM process. Taken from IBM's knowledge center.

Data reduction strategies can be divided in dimensional reduction ones where it is attempted to project the data into a smaller set, numerical reduction is based on using smaller forms of data representation and compression where a smaller data set is constructed with no information loss.

In the context of work done this step will encompass the transformation of the drugs into molecular descriptors and the treatment of the cell line features.

Feature Selection

Feature selection in data mining is the process of selection a subsection of the feature space with minimal information loss. This is accomplished by finding and removing present mostly redundant data [33]. There are various benefits in having a smaller feature subset [34]. The risk of over-fitting the model to the training will be smaller. Data will be easier to interpret and visualize and the model will be trained faster using less storage. The identification of redundant data is also

valuable in itself as it allows for better decision-making regarding either the extraction or addition of new features.

The usual goal with feature selection is to find the *minimal-optimal* set of features, the feature set that obtains the most relevant (optimal) classifier using the smallest possible number of features [35]. Due to the importance of finding every feature related to the target variable, in some areas, such as bioinformatics the goal is instead to find all of these relevant features. This constitutes the *all-relevant* problem. Current literature focuses on these two formulations and in this thesis both will be explored.

Feature selection is a complex problem as variables that have no information by themselves can be relevant together and highly correlated variables still add information [34] and thus simply ranking the feature may prove insufficient. Methods are commonly grouped by their learning process into wrappers, filters and embedded methods which generate subsets and evaluate them using machine learning models, filters, which perform the selection process before the model is run and embedded methods which occur during the training process of a machine learning model [34].

In the type of data which will be used in this thesis, genomic data, feature selection is a particularly relevant area [34]. Since genomic datasets often consist of numerous features with many of them redundant there is a wealth of knowledge to draw upon and there are methods specifically implemented to solve problems in this area.

Boruta

Boruta is a feature selection algorithm based on the feature relevance comparison between the original features and random added ones [36]. Random features are copied from existing ones and then shuffled so that there is no correlation to the target variable. Afterwards the features are ranked according to their gini impurity. Features score significantly higher than average variable importance rating of the added random variables, plus a customizable threshold, are deemed important and features that score significantly lower are discarded. The rest of the attributes are considered to have undetermined importance and move to a new round. The process is repeated until all attributes' importance is determined or a set limit of rounds is reached.

Dimensionality Reduction

Dimensionality reduction methods condense the same amount of information in a smaller dataset. They differ with feature selection methods in that feature selection operates on feature space by removing or adding them but don't remove features outright.

Principal Component Analysis

Principal Component Analysis (PCA) is the most common dimensionality reduction method and it is frequently used in machine learning contexts **anton** PCA computes an orthogonal transformation of the original features into a using a different smaller set, called principal components,

in a different coordinate system. This is done through the computation of the eigenvectors of the feature spaces' covariance matrix. The number of eigenvectors chosen determines the number of new features, it is common to try to retain around 99% variance.

Machine Learning

Machine Learning (ML) algorithms will be used for the data mining section of the knowledge discovery process. ML consists of allowing a machine to acquire information without being explicitly programmed it allows for pattern recognition on large amounts through the use of a training set with the results previously labeled or by data exploration if there is no training set. ML can be used in a supervised learning context where we use a dataset with pre-calculated results to train the model or unsupervised learning where the data is unlabeled. Supervised learning ML algorithms deal with regression and classification problems. In regression the value of a continuous variable is predicted whereas in classification problems we try to group objects into classes. In unsupervised learning the most common problem is clustering that consists of grouping up similar objects. In the following sections the algorithms planned to be used will be described.

Support Vector Machines

Support vector Machines (SVM) [37] are a supervised learning method that can be used for either classification and regression. The model attempts to find the hyperplanes that best separate the data. In order to be applied to cases where the data is not linearly separable the features of the data must be transformed into a higher space however in this higher dimensional feature space the calculations would be too computationally expensive. To solve this problem functions that approximate the original calculations called kernel functions are often used. There are several different types of kernel functions, the choice of which to use depends on the problem and is often only found through experimentation. The versatility allowed by the different kernel functions and the ability to deal with a high number of features make SVMs a commonly used machine learning model.

Artificial Neural Networks

Artificial Neural Networks (ANN) [38] consist of an attempt at the emulation of the biological networks of neurons present in the brain. The neural network consists of a set of parallel units, usually called layers, that respond to a set of given inputs. Each unit in a layer receives an input or a set of inputs and processes them into an output which is fed into a set of units in the next layer. A neural network consists of an input and an output layer, with an optional amount of layers in between called the hidden layers. A network that has at least one hidden layer is called a multilayer network [39]. The additional layers allow these networks to represent a wider range problems besides linear ones. The function that processes the inputs of each unit the output is called the activation function. There are many different activation functions, the more common ones are the sigmoid function and the hyperbolic tangent.

Supervised learning in neural networks is done by progressively updating the weights using the difference of the output and the target multiplied by a user defined learning rate. In multilayer networks a backpropagation method is commonly used. This method calculates the difference between the output and the goal at every unit and calculates the new weight using a percentage of gradient descent.

Decision Trees

Decision Trees (DT) [40] are a family of algorithms that classify the data through tree-like structures. The data is divided at decision points in the structure and leaf nodes represent the final segments of the data. Decision trees are constructed by recursively using a greedy strategy that divides the data according to a metric that varies between algorithms until a stopping criteria is met. The most popular algorithms that use decision trees are C4.5 [41] and CART [42]. C4.5 is an evolution of the ID3 algorithm and uses entropy as the metric to minimize at decision points. CART uses Gini impurity instead and allows for regression along with classification.
Decision Trees have the main advantage of being a white-box model, which means we can interpret the decisions made by the algorithm. Another of the advantages is that feature are individually analyzed, this makes decision trees robust to non-centered and disparate features. In spite of this, they are seldom used due to their high variance.

Ensemble Methods

Ensemble methods [43] combine various types of models in order to achieve better performance. The use of an aggregate of multiple classifiers is done the robustness over a single instance of the algorithm. To combine the various instances ensemble methods either average their predictions or use the instances sequentially. To introduce randomness random subsets of the feature space and are distributed by the various instances used, this process is used on some ensemble methods and is called bagging. Some examples of ensemble methods are Adaboost [44] and Random Forests [45].

Random Forests

Random Forests [45] are an ensemble method that uses decision trees as a base learner using feature bagging. Each tree then votes for a model and the most popular one is chosen. Random Forests can be used both for regression and classification. They retain the robustness to different types of features from decision tress while removing the variance issues. This, with the additional factor that random forests produce interpretable results, makes them a commonly used algorithm in the context of cheminformatics. There are metrics associated with Random Forests that are associated with feature importance. Gini impurity, used in classification trees, measures how often a feature would correctly classify a random item. In regression problems the Residual Sum

of Squares, which measures the difference between predicted and actual values is used to perform splits and the mean squared error is used to measure feature importance instead.

Relational Algorithms

Relational machine learning algorithms are algorithms that take structured data such as graphs or xml as input instead of a single data table. In this section two of the main approaches will be explored.

Graph Mining

Graph Mining algorithms are a family of algorithms that try to find common substructures in a graph. As molecules can be represented as graphs graph mining is often used in chemistry related scenarios and because of that this family is of particular relevance to this project. The process of graph mining can be divided in two main steps. The first step consists of finding the sub-graphs to test for frequency and the second is the test of the frequency of the sub-graphs chosen.

Frequent Subgraph Mining

Frequent Subgraph Mining (FSM) attempts to find frequent graph structures in one or more graphs [46]. Algorithms are categorized into either pattern growth algorithms or apriori-based. Apriori algorithms are simpler but less efficient, they rely on joining graphs, starting with a single vertex, and testing whether or not is it is frequent. Pattern growth approaches rely on extending an existing subgraph through edge addition. The most common applications are in bioinformatics as molecules can easily be represented in graphs.

gSpan

gSpan is a frequent subgraph mining algorithm that follows a pattern growth approach [47]. The gSpan algorithm starts by creating a unique representation of a graph, called a DFS code. First each edge is assigned a code with the following format (x,y,z,i,j) where x is the starting vertex identifier, y the ending vertex identifier, z the vertex type, i the edge type and j the ending vertex type. Then edges are ordered through a number of rules based on the assigned identifiers at the start, the basic principle is that an edge connecting 2 vertexes cannot have the starting vertex higher than the ending one with some exceptions for backwards edges and edge cases. To guarantee uniqueness possible graph DFS codes can be compared and ordered and only the minimum set is chosen. The ranking is done through a set of eleven rules with the general objective of minimizing backwards edges. The second phase of the algorithm then expands from edges found in all the graphs where the count is higher than the support level, which is the given frequency number for which we are trying to find common subgraphs. It then extends the edge using a depth first search first trying

to connect existing vertexes and after expanding the graph outwards checking the support level and discarding if it's not common enough. As the candidate graph is generated every subgraph is pruned from future exploration or extension, since the search is depth-first a significant amount of candidates can be pruned. An example of the output of the gSpan algorithm can be seen on Listing 2.1, an example of the input would simple the vertexes and edges, without the first and last line.

Listing 2.1: gSpan output example - the first line contains the id (8) and the number of times the subgraph has been found. The following lines either vertexes following the 'v x y' format x is the id and y the atom type or the edges using the format 'e x y z' where x is the starting vertex id, y the ending vertex id and z the connection type. The last line contains references to the graph ID's in the input file where this subgraph was found

```
t # 8 * 23
v 0 6
v 1 6
v 2 6
v 3 6
v 4 6
v 5 6
v 6 6
v 7 6
v 8 6
e 0 1 1
e 1 2 1
e 2 3 1
e 3 4 1
e 4 5 1
e 5 6 1
e 6 7 2
e 7 8 1
x 0 3 5 12 16 18 29 37 41 42 43 44 46 53 68 78 79 89 97 100 101 103 104
```

Inductive Logic Programming

Inductive Logic Programming (ILP) [48] combines machine learning with logic programming. ILP is a supervised learning method that requires that the background knowledge be encoded as first-order logic rules and that a set of examples, either positive or negative, be given also encoded as logic rules. ILP, as its name indicates uses inductive reasoning which generalizes known examples to reach an hypothesis. An ILP algorithm usually consists of two phases. The first is the generation of the candidate hypothesis and the second is the verification of the hypothesis generated.

ILP has been used with success in chemical problems [49]. The structure of a molecule can easily be represented in first order logic rules with the ability to represent the connections between atoms being of particular relevance. Another strength is the fact that ILP models output easily readable rules.

Model Evaluation

Evaluating a model is an important part of the knowledge discovery process, it is necessary to know what needs correcting and what the next steps should be. In order to avoid bias it is important to construct separated training and testing sets. One common practice is to divide the labeled dataset into a training set and a test set, this allows for the calculation of an error rate through the prediction of the test set elements in the model previously trained in the training set. This simple method is named Hold-Out validation. K-fold Cross Validation is a similar technique that divides the dataset into k sets and each one of those sets take a turn being the test set while the other k-1 sets are used as the training set.

Metrics

There are several metrics used in machine learning, with different ones being used in classification or regression problems [50].
In classification metrics are based on the accuracy of the model. The simplest measure available is accuracy, which is simple the division of the correct predictions by the number of cases in the dataset. Precision and recall are two metrics that are often used together in classification problems; recall 2.1 measures the amount of positive examples out of all possible examples found and precision 2.2 is the percentage of positive examples marked as such out of all positive examples in the dataset. The F1 score 2.3 can be explained as a measure that takes both precision and recall into consideration and performs an harmonic mean.

$$Recall = \frac{PositiveExamplesIdentified}{AllExamples} \tag{2.1}$$

$$Precision = \frac{PositiveExamplesIdentified}{AllPositiveExamples} \tag{2.2}$$

$$F1 = 2\frac{precision.recall}{precision + recall} \tag{2.3}$$

Another more complex metric commonly used is log-loss which requires the probability output of each classification and uses it to measure the model as described in equation 2.4.

$$LogLoss = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j}\log(p_{i,j}) \tag{2.4}$$

In regression the basic measurement is the sum of the differences between the correct values and the predicted values. The most popular metric is the root mean square error which is described by

the formula 2.5.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(c_i - \bar{c}_i)^2} \tag{2.5}$$

Another measure used is the median absolute percentage (MAPE) 2.6 which uses the median to avoid the effect of outliers.

$$MAPE = median(|(c_i - \bar{c}_i)/c_i|) \tag{2.6}$$

Also used in sciences is the Pearson correlation coefficient which measures correlation with 1 meaning total linear correlation, 0 no correlation and -1 negative correlation. With x and y meaning actual and predicted values the formula is as seen in equation 2.7.

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.7}$$

Tools

There are many free and efficient implementations of the above machine learning algorithms. Weka [51] is a software package developed in java by the University of Waikiti from New Zealand. Weka contains pre-processing algorithms and all of the non relational algorithms mentioned above. It is freely available for non commercial projects under the GPL 3.0 license.

Scikit-learn [52] is a python library that also provides the most common machine learning algorithm. It is build on Python with Cython being used for some algorithms for efficiency reasons. It is free to use under the BSD license. R [53] is programming language designed for statistical problems. Machine Learning algorithms are available in the form of packages. Since its development it has gained popularity as a data mining tool due to wide ranged of available algorithms and its ease of use. R is freely available for non commercial products under the GNU General Public license. The gSpan algorithm is available as an executable developed by the author at https://www.cs.ucsb.edu/ xyan/software/gSpan.htm.

2.4 Related Work

Menden *et al.* [5] developed the first solutions for the prediction of IC50 values. Models were built using as inputs biomarkers and copy number values from the cell tissue and 1 and 2 dimensional descriptors with molecular fingerprints from drugs. Two dataset 8 fold splits were done, one randomly and the other with no cell lines repeated acrros folds. Two models were then created for each dataset, one with random forests and another with SVM's both achieving 0.82 RMSE predicting the IC50 values on the test fold having been trained on the rest of the folds and 0.96 on a blind test set. This was a 20% increase in performance over non-machine learning methods.

In João Ladeiras' dissertation [6] a similar cancel drug effectiveness model was built using machine learning methods. The model was also based on a replication of Menden *et al.* and was transformed into a classification problem and expanded with ILP and FSM. The replication results were similar with 0.83 RMSE in the test fold and 0.93 in the test set. The models used were

SVM's, Random Forests and Neural Networks. After this parameter tuning was added having a 1% increase. After that a selection of gSpan features was added with no significant performance increase. The problem transformed into a classification one and ILP also with no significant performance increase. Costello et Al.[54] tested various algorithms for cancer drug prediction on an NCI60 based dataset concluding that a novel algorithm, Bayesian multitask MKL, achieved the highest performance. The evaluation metric used was a custom variant of the concordance index, a common metric in survival analysis.

Iorio *et al.* [55] produced a model using the GDSC dataset and tumour derived cell lines from eleven thousand patients. The paper introduced cancer functional events as predictors which were defined as a collection of features that correlated positively with cancer existence in a cell lines. Experiments were conducted using analysis of variance, machine learning algorithms and a novel logical model named *Logic optimization for binary input to continuous output (LOBICO)* arriving at a median Pearson correlation of 0.92 in pan-cancer models using as inputs cancer genes, focal recurrently aberrant copy number segments and hypermethylated informative 5'C-phosphate-G-3' sites in gene promoters. Machine learning models were also used to compare feature importances with the conclusion that genomic features were the most important towards IC50 prediction.

Chapter 3

The methodology

In this chapter the methods and approach that were followed in the experiments are described. First the dataset preprocessing and replication of Menden *et al.* [5] is detailed and following that the experiments, including methods and reasoning. The experiments that were performed start with the testing how the model responds to different amounts of data, following that the feature space of the replication model is analyzed and the problem is transformed into a classification one. The final experiment described is the addition of drug features obtained with Graph Mining to the feature space.

3.1 Original Data

The base dataset that was utilized in our project is the 1.0 version of the Genomics of Drug Sensitivity in Cancer (GDSC) project [56] [1]. The dataset contains IC50 values, cell line characterization and other information for 638 cell lines and 136 drugs. Only 58% of drug-cell line pairs have IC50 values with remaining pairs being blank. The 1.1 version added 15,578 cell/drug pairs, these were used to build an additional blind test set.

In addition to the base GDSC data, generated drug features, obtained using the PaDEL software, were also added to the dataset. These features will be detailed in the following sections.

3.1.1 Genomic Features

Each cell line is characterized by 155 genomic features – the genetic mutation data consists of the Microsatellite instability (MSI) status, the coding variant and copy number variation for 68 tissue biomarkers making a total of 137 features. The features are encoded in the dataset as follows:

- MSI status is 1 if unstable and 0 if stable.
- The phenotype and copy number are put together in an object with the "X::Y" form where X is the phenotype and Y the copy number.

[1] Available at www.cancerrxgene.com

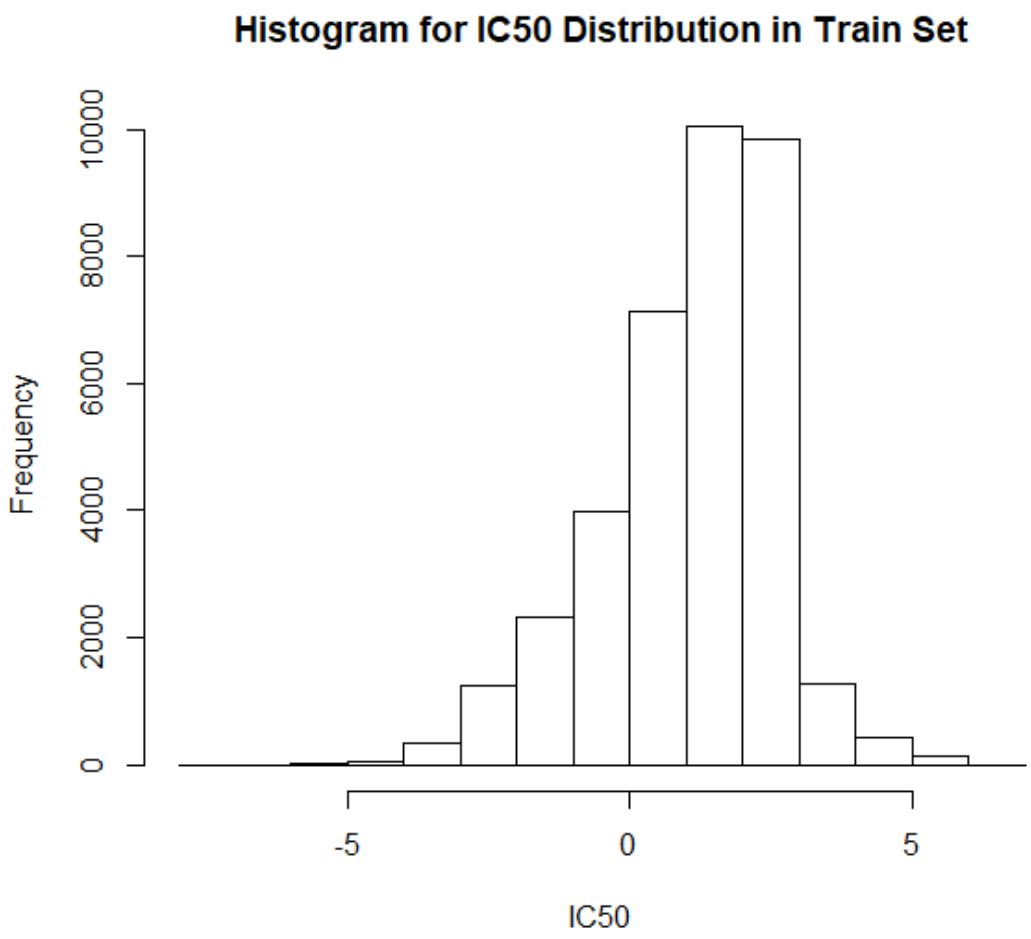

Figure 3.1: IC50 Values Histogram

- The coding variant is either "wt" (wildtype) or a mutation which starts with "p."
- The copy number is either ">=8" (amplification), "0" (deletion) or "0<cn<8" (wildtype)

To simplify the features for machine learning purposes, the copy number and coding variant were transformed into numerals through the following rules:

- The coding variant is 0 if it's wildtype or 1 if it's a mutation.
- The copy number is 0 if it's wildtype, 1 if an amplification and -1 if it's a deletion.

3.1.2 Drug Features

For the 110 drugs for which there were a SMiLEs string available, it was obtained using the PubChem API. Using the SMiLES string and PaDEL version 2.11 the 1D and 2D descriptors and molecular fingerprints were obtained. Some examples of features are the number of times each

atom is found, topological graph indexes (Zagreb Index) and for the fingerprints - the existence or not of an N-N bond. The total number of drug features is 1603 with 722 being 1D and 2D descriptors and 881 molecular fingerprints.

3.2 Data Preprocessing

The steps taken by Menden *et al.* [5] were used as a reference as those are the results we replicate as a basis. The data was organized as cell line/drug pairs each having an IC50 value, pairs that did not have an IC50 were dropped leaving a total of 38,654 pairs left.

Genomic features with more than 15 missing genomic values were removed, which resulted in the deletion of 61 cell lines. Every feature with missing values was removed. For the drug features all of those with missing values were removed. The final dataset consists of 790 total features of 38654 drug/cell pairs and 15578 extra testing pairs. The full list of features is available on appendix A.

The results of the preprocessing were similar to Menden *et al.* but not equal, the final result for the preprocessing was 790 features in our case and in the original paper it was 827. These results can be explained by difference in software versions such as PADeL or unclear procedures in the original paper, a full list of features is not provided so there can be no confirmation of what extra features were removed. Ladeiras arrived at the number of 932 features, which further corroborates the unclarity of the procedures in Menden *et al.* For the rest of this document the test fold will be the left-out fold in cross validation, the train set the folds used to train the model and the test set the 15,768 extra pairs in the 1.1 version.

3.3 Replication of Menden *et al.* random forests experiment

The first experiment is replicating Mendel et Al's [5] random forests model in order to use the results as a baseline. As in the original experiment, the dataset is separated into 8 folds with 1 left out for testing and with no parameter training performed. 2000 trees were used, since no number is given. The relatively high number serves to err on the side of caution. The model is built using the R language with the 'randomForest' package using 2000 decision trees. The results can be seen in the next chapter in Table 4.1.

3.4 Experiments

The methodology for the experiments performed after the replication of the original study are described in the following sections. A significant body of literature [57]–[59] provides solid evidence that improving the feature space is more relevant to better model creation than the algorithm used. This influenced the choice of the experiments done, no changes to the algorithm were made and the focus was on adding or improving existing features. At the start we performed a learning curve and analyze how the model performs under different amounts of data, after that the performance

of the different types of features are analyzed and finally we explore the addition of Graph Mining features. The results are presented and discussed in Chapter 4.

3.5 Learning curve analysis

In order to better understand the behavior of the model when given different amounts of data, the previous experiment was repeated while withholding certain data percentages on the left-out fold and on the 1.1 test set. Also added was the result of evaluating the model on the same data it was trained on in order to evaluate over-fitting. We fed the model different amounts of data in steps of 20%, ending with the percentages of 20, 40, 60, 80. We also add 2% to this list to have an initial low value. Understanding this behavior is highly relevant in the context of cell line experiments, as those experiments are cumbersome to perform, therefore evaluating whether more examples correlate with a a better model using the same features is valuable. In addition, despite the simplicity of the learning curve, the method has never been applied regarding this dataset which also adds value to this approach.

3.6 Feature Selection

Understanding how the features affect a machine learning model is key for it's improvement. There are several insights that can be gained from this – which features provide the most information gain between genomic cell line information and drug features and also among their subcategories. This can enable the removal of irrelevant features or to direct research and improvements towards them. In order to do this metrics produced by the random forests will be analyzed, these measure the averaged MSE and node purity, the first is a feature importance rating following the methodology for these in the previous chapter, the second indicates how well the data is split by that feature. These are included in the *randomForests* R package which was used to build the model. In addition the Boruta [36] feature selection algorithm was ran to compare features against random noise and classify them either useful or not. A limit of 100 rounds, the default value, was imposed on Boruta due to time constraints.

3.7 Classification

After the regression analysis the problem was transformed into a classification one so as to explore it's performance in a different type of model, if it's easier to classify cell-drug pairs as having a strong reaction or not and if those categories provide an accurate enough reaction description. The categorization of drug-cell IC50 reaction are as follows:

- Lower than 1.3 IC50 was coded as -1
- Between 1.3 and 1.5 IC50 was coded as 0
- Higher than 1.5 IC50 was coded as 1

This is based on the IC50 histogram in Figure 3.1. After this the first random forests experiment was repeated using 128 trees and 4 folds. The reduced number of folds and trees is due to time constraints.

3.8 Frequent Subgraph Mining

Graph Mining was used to help in unveiling common substructures of the applied drug that could help domain experts explain it's effect on the cell lines. The possibility of characterizing the drugs with their molecular substructures, adding them to the remaining features, was explored. In order to find the substructures the drugs were represented as graphs and Frequent Subgraph Mining algorithms were ran. This shares some similarities with the molecular fingerprints, the goal is to find out whether by creating a tailored feature set for the drugs present in the dataset the fingerprints can be outperformed and also to evaluate the performance of the approach in general.

The algorithm chosen was for Frequent Subgraph Mining was gSpan [47], due to it's performance and availability. The implementation used was the 64-bit original executable [2]. The algorithm was ran with a support level of 20%. The FSM ecosystem does not have readily available converters so custom solutions were used. To convert the data from the sdf format to gSpan a tool by Andreas Maunz[60] was used. GSpan results were then converted to per graph columns were 1 indicates the presence of the subgraph in the drug and 0 the fact that it is not present. This resulted in the addition of 919 features top to the dataset resulting in 1710 total features.

Two experiments were ran, first a replication of the first one with 2000 trees and 8 folds and another one where the features were considered factors with a reduced number of trees – 64 and only 4 folds; This is because with the full gSpan dataset each tree takes around an hour to be processed. The low number of trees may have some impact on performance but Ladeiras ran similar experiments with lower tree numbers with no significant performance reduction.

[2] Available at *https://www.cs.ucsb.edu/ xyan/software/gSpan.htm*

Chapter 4

Discussion of the Results

In this chapter we present the results, starting with the replication of the Menden *et al.* 's experiment [5] and moving on to the learning curve, feature selection and FSM features discussing the results in the context of the problem.

4.1 Replication of the original study

The results indicate a successful replication of Menden *et al.* 's experiment, the RMSE was 0.82 in the validation fold while our result was 0.83. The small difference can be explained by the difference in features described in the previous section. The RMSE on the blind test set was 0.98 also similar to the 0.97 in Menden *et al.*, about 5% higher than the result Ladeiras' model [6] had in the same experiment (0.92), this can be explained by the differences in data preprocessing and in decision tree number. It can be concluded that the results are similar, even tough there were some discrepancies in the feature sets, the similarity in results allows the model to be used as a baseline for future experiments.

Table 4.1: Per fold results of the replication of Menden *et al.* random forests with the Root Mean Square Error (RMSE) and Pearson Correlation (R)

Fold	RMSE	R
1	0.833	0.71
2	0.817	0.719
3	0.833	0.702
4	0.826	0.722
5	0.825	0.71
6	0.853	0.691
7	0.848	0.696
8	0.823	0.714
Mean	0.83225	0.708
Std Dev	0.0125	0.109

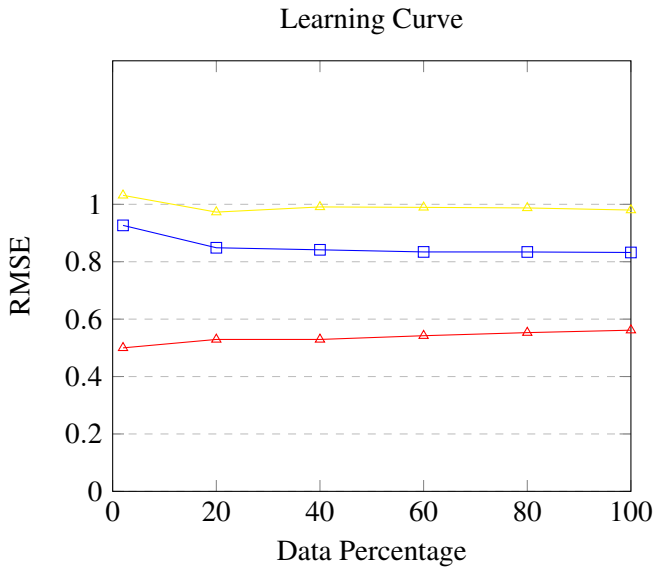

Figure 4.1: Line plot of the results of the prediction on the left-out validation fold (blue), the version 1.1 test set (yellow) and train set (red)

4.2 Learning Curve

The results of the learning curve, in which the model was fed different amounts of data, can be seen on Figure 4.1 and the full results are available on Appendix B.

The main takeaway is that no significant gains are had with higher amounts of data, even with 20% of the data used the results are similar to when the full dataset is used. This indicates that priority should be given to either improving the models or the feature set. The performance of the model is higher in the train set which can indicate overfitting. Further corroborating this is the fact that the blind test set, which contains different cell line pairs has worse performance. Despite this when more data is added there is no performance loss, which shows that it's not a problematic case.

4.3 Feature Selection

The results of the Boruta algorithm regarding confirmed or rejected features can be seen on Table 4.2 with the full results available on Appendix C. From the 790 total features 178 were confirmed, 446 rejected and 176 were left uncategorized due to the 100 round run limit.

Table 4.2: Boruta rejected and confirmed features

	Genomic	Molecular 1 and 2D	Molecular Fingerprints	Total
Rejected	91	25	329	445
Confirmed	13	126	39	178
Tentative	2	4	161	167
Total	106	155	529	790

Even though the 100 run limit was imposed, most features were categorized by the Boruta algorithm, while taking 1 week to run on a 4 core modern CPU computer, therefore we can conclude that the round limit was a reasonable choice. The first conclusion drawn from the results is that a significant amount of features are low quality and perform no better than random noise. Molecular 1 and 2D descriptors performed consistently with almost all of them being considered relevant. The molecular fingerprints and Genomic features performed rather poorly with only 7% and 13%, respectively, being considered relevant.

To complete our feature analysis we can also see the results of the built in average MSE feature ranking and node purity by the *randomforest* package on Figure 4.2. These results indicate the The full importance results for one fold can be visualized on Appendix E. The results shown are for one fold only, but all of the folds exhibit the same behaviour. It is possible to see that the features with the most information gain were genomic ones. This can be explained by the fact that this is a pan-cancer model, and the model is more tuned to a particular type of cancer the features for which have a high amount of information, while other similarly typed cell line features end up having lower contributions. There is no information about cancer type in this version of the dataset. This is consistent with Iorio *et al.* where cancer specific models had better performance. Molecular 1D and 2D features provide consistent information gain but do not rank very high, showing they are consistent but not overly important. Even though the feature sets were different, especially in the cell line information, gene expression features, represented here by the phenotype (wildtype), were the most important features in an experiment conducted with Random Forests by Iorio *et al.* which is also consistent with the findings here.

4.4 Classification

The results of the averaged 4-fold classification can be seen on Tables 4.3, 4.4 and 4.5. The results of the 0 class were omitted since its prevalence was around 5% which makes the results too disparate. In hindsight it would have been preferable to either remove the class or expand the 1 or -1 classes.

It can be concluded that the results prove that this transformation is at least viable with good accuracy results. Comparing to Ladeiras our results show worse performance in precision and recall, for example in validation fold precision and recall are 0.9 in Ladeiras' version while we achieved 0.8 and 0.7, respectively. The difference in features - described in the replication results can account for some of the differences with the rest being due to a difference class split.

Table 4.3: Classification accuracy on the validation fold and set

	validation fold	Test Set
Accuracy	0.7836	0.769

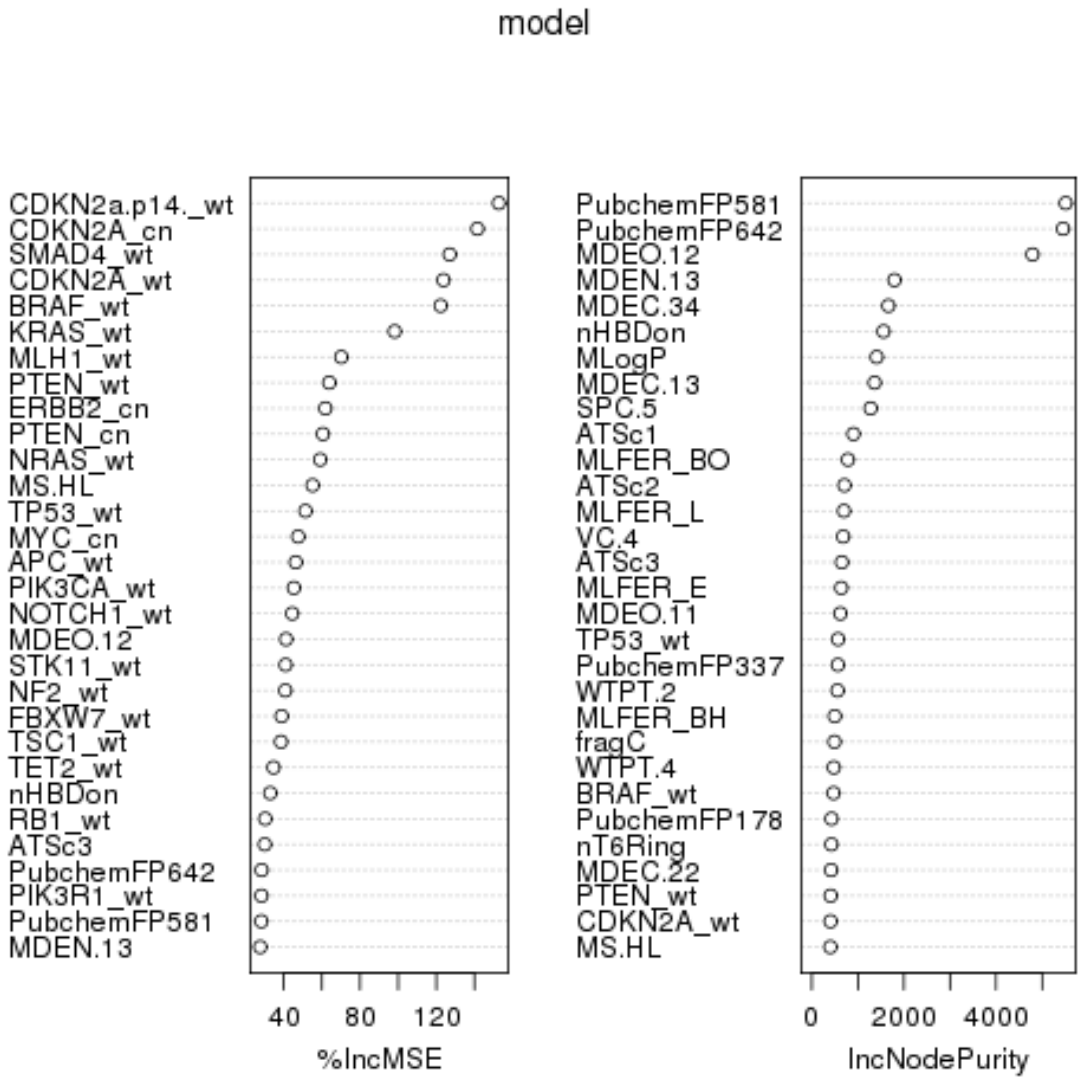

Figure 4.2: Random Forest importance for the first fold in the first experiment

4.5 gSpan

The random forest experiment was reran with the same parameters as the first experiment with the results seen on Table 4.6 and also with gSpan features as a factor in Table 4.7. The results on the test set were the same to two decimal places – 0.98 RMSE. There was no significant improvement and the results are mostly similar.

Through the analysis of the importance scores produced by the random forests model, from which the top 30 features can be seen in Figure 4.3 with the rest available in Appendix D, it is indeed possible to see that the average MSE of these features is quite low – 3.3, with the genomic features having, as an example, 22 average MSE even though some of them are negative and the number 1 ranked feature having 120 average MSE. The molecular fingerprints features had an

Table 4.4: Classification values on class 1 and -1 on the validation fold

	Precision	Recall	F1
1	0.83	0.78	0.80
-1	0.82	0.769	0.79

average MSE of 3.12, which means that the added gSpan features had a higher overall importance. In short the prevalence of the cell line features has eclipsed the results of other features, including the gSpan ones that were added. This is also consistent with the results obtained by Ladeiras in a similar experiment also with gSpan features, though they were heavily preprocessed.

4.6 Final result analysis

Reviewing all the tests there are significant conclusions to be had, the first ones relate to the model itself; The learning curve shows the model doesn't improve with more data and the Boruta algorithm and feature importance ratings show that a significant number of features are not superior to noise. This shows that there is room for improvement in the original model, improving the features may resolve the behavior seen in the learning curve. Drug features, though more numerous contributed significantly less to the model than cell line ones, which indicates that the addition of more characterization of the cell lines could yield better results. Recent research has been in this direction[55] further corroborating the conclusion.

Regarding gSpan features, they are still an interesting exploration pursuit even though the results weren't particularly encouraging. The experiment performed allowed for an evaluation of the importance of most FSM features since no features were removed. In the end they ended being drowned by other features, particularly genomic ones.

Table 4.5: Classification values on class 1 and -1 on the test set

	Precision	Recall	F1
1	0.85	0.77	0.81
-1	0.79	0.75	0.75

Table 4.6: Per fold results of the replication of Menden *et al.* random forests with the addition of gSpan features with the Root Mean Square Error and Pearson Correlation (R)

Fold	RMSE	R
1	0.833	0.71
2	0.817	0.719
3	0.833	0.702
4	0.826	0.722
5	0.825	0.71
6	0.853	0.691
7	0.848	0.696
8	0.822	0.713
Mean	0.8321	0.708
Std Dev	0.0125	0.011

Table 4.7: Per fold results of the replication of Menden *et al.* random forests with added gSpan features as a class with the Root Mean Square Error and Pearson Correlation (R)

Fold	RMSE	R
1	0.83	0.71
2	0.83	0.72
3	0.84	0.71
4	0.85	0.72
Mean	0.84	0.71
Std Dev	0,01	0,006

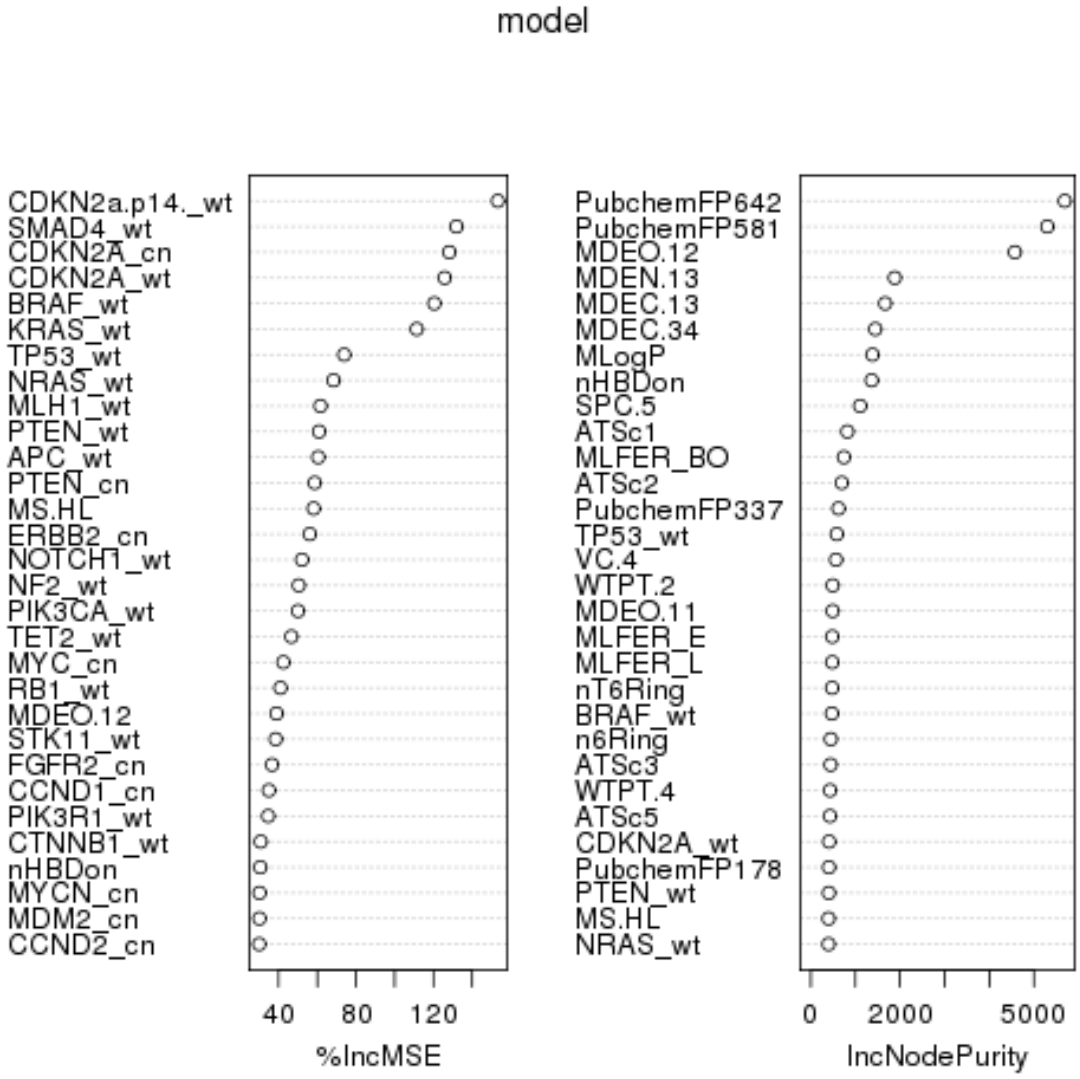

Figure 4.3: Random Forest importance for the first fold in the gSpan experiment

Chapter 5

Conclusions and Future Work

In this chapter the overall goal satisfaction and future work possibilities are explored.

5.1 Goal Satisfaction

The goal of gaining insight and providing solutions to improve model performance was mostly achieved. The learning curve will help inform future decisions regarding dataset expansion and the application of frequent subgraph mining in this problem in this form was novel in spite of non-stellar results. The main limitation factor throughout this project was the time required for the model to be trained on the available machines, nevertheless the analysis that was performed is valuable for future approaches to the problem and can be generalized to future approaches in this problem. Perhaps the biggest shortcoming in the project was sticking too closely to Menden *et al.* 's model, developing different metrics and using different datasets so as to able to better compare against other studies.

5.2 Future Work

Since the run time of the experiments was quite high some possibilities were not explored and thus there are a number of possible paths regarding future work. Regarding the gSpan experiments there are interesting possibilities to try. Removing the molecular fingerprints after adding the gSpan features could produce better results since there is some redundancy regarding fingerprints and gSpan features. Another possibility is removing all gSpan features that are present in all drugs under a certain threshold. Those molecular substructures may be too common too characterize a drug and may be only adding noise reducing overall performance. Experimenting with different support levels in the gSpan algorithm or trying different FSM algorithms altogether would also be a worthwhile pursuit. Exploring the use of deep learning in this particular problem

would be interesting since there is literature that indicates performance gains in a cheminformatics context[61].

Appendix A

Features

Feature Name	Feature Name	Feature Name	Feature Name
MS-HL	AKT2_cn	ALK_wt	APC_wt
APC_cn	BRAF_wt	BRAF_cn	BRCA1_wt
BRCA2_wt	BRCA2_cn	CCND1_cn	CCND2_cn
CCND3_cn	CDH1_wt	CDH1_cn	CDK4_cn
CDK6_cn	CDKN2A_wt	CDKN2A_cn	CDKN2C_wt
CDKN2C_cn	CDKN2a(p14)_wt	CTNNB1_wt	CTNNB1_cn
CYLD_wt	EGFR_wt	EGFR_cn	EP300_cn
ERBB2_wt	ERBB2_cn	EZH2_wt	EZH2_cn
FAM123B_wt	FAM123B_cn	FBXW7_wt	FBXW7_cn
FGFR2_cn	FGFR3_wt	FGFR3_cn	FLCN_wt
FLT3_wt	FLT3_cn	GNAS_wt	GNAS_cn
HRAS_wt	IDH1_wt	IDH1_cn	JAK2_wt
JAK2_cn	KDM5C_wt	KDM5C_cn	KDM6A_wt
KDM6A_cn	KDR_cn	KIT_cn	KRAS_wt
KRAS_cn	MAP2K4_wt	MAP2K4_cn	MDM2_cn
MET_cn	MLH1_wt	MLH1_cn	MLLT3_cn
MSH2_wt	MSH2_cn	MSH6_wt	MSH6_cn
MYC_cn	MYCL1_cn	MYCN_cn	NF1_wt
NF1_cn	NF2_wt	NF2_cn	NOTCH1_wt
NRAS_wt	NRAS_cn	PDGFRA_cn	PIK3CA_wt
PIK3CA_cn	PIK3R1_wt	PIK3R1_cn	PTEN_wt
PTEN_cn	RB1_wt	RB1_cn	RUNX1_wt
SETD2_wt	SMAD4_wt	SMAD4_cn	SMARCA4_wt
SMARCA4_cn	SMARCB1_cn	SMO_cn	SOCS1_cn
STK11_wt	STK11_cn	TET2_wt	TP53_wt
TP53_cn	TSC1_wt	TSC1_cn	VHL_wt
VHL_cn	WT1_cn	PubchemFP0	PubchemFP1
PubchemFP2	PubchemFP6	PubchemFP9	PubchemFP10
PubchemFP11	PubchemFP12	PubchemFP13	PubchemFP14
PubchemFP15	PubchemFP16	PubchemFP17	PubchemFP18
PubchemFP19	PubchemFP20	PubchemFP21	PubchemFP22
PubchemFP23	PubchemFP24	PubchemFP25	PubchemFP30

PubchemFP33	PubchemFP34	PubchemFP37	PubchemFP38
PubchemFP43	PubchemFP44	PubchemFP46	PubchemFP93
PubchemFP115	PubchemFP116	PubchemFP117	PubchemFP118
PubchemFP129	PubchemFP130	PubchemFP132	PubchemFP143
PubchemFP144	PubchemFP145	PubchemFP146	PubchemFP147
PubchemFP148	PubchemFP149	PubchemFP150	PubchemFP152
PubchemFP153	PubchemFP155	PubchemFP156	PubchemFP157
PubchemFP159	PubchemFP160	PubchemFP164	PubchemFP167
PubchemFP178	PubchemFP179	PubchemFP180	PubchemFP181
PubchemFP182	PubchemFP183	PubchemFP184	PubchemFP185
PubchemFP186	PubchemFP187	PubchemFP188	PubchemFP189
PubchemFP190	PubchemFP191	PubchemFP192	PubchemFP193
PubchemFP194	PubchemFP195	PubchemFP199	PubchemFP200
PubchemFP206	PubchemFP213	PubchemFP214	PubchemFP218
PubchemFP219	PubchemFP227	PubchemFP228	PubchemFP232
PubchemFP233	PubchemFP241	PubchemFP246	PubchemFP247
PubchemFP248	PubchemFP252	PubchemFP255	PubchemFP256
PubchemFP257	PubchemFP258	PubchemFP259	PubchemFP260
PubchemFP261	PubchemFP262	PubchemFP274	PubchemFP294
PubchemFP297	PubchemFP298	PubchemFP299	PubchemFP300
PubchemFP301	PubchemFP305	PubchemFP308	PubchemFP314
PubchemFP327	PubchemFP328	PubchemFP330	PubchemFP332
PubchemFP333	PubchemFP334	PubchemFP335	PubchemFP336
PubchemFP337	PubchemFP338	PubchemFP339	PubchemFP340
PubchemFP341	PubchemFP342	PubchemFP344	PubchemFP345
PubchemFP346	PubchemFP347	PubchemFP349	PubchemFP350
PubchemFP351	PubchemFP352	PubchemFP353	PubchemFP355
PubchemFP356	PubchemFP357	PubchemFP358	PubchemFP359
PubchemFP360	PubchemFP362	PubchemFP363	PubchemFP364
PubchemFP365	PubchemFP366	PubchemFP367	PubchemFP368
PubchemFP370	PubchemFP371	PubchemFP372	PubchemFP373
PubchemFP374	PubchemFP375	PubchemFP376	PubchemFP377
PubchemFP378	PubchemFP379	PubchemFP380	PubchemFP381
PubchemFP382	PubchemFP383	PubchemFP384	PubchemFP385
PubchemFP386	PubchemFP387	PubchemFP388	PubchemFP389
PubchemFP390	PubchemFP391	PubchemFP392	PubchemFP393
PubchemFP394	PubchemFP395	PubchemFP396	PubchemFP397
PubchemFP398	PubchemFP399	PubchemFP400	PubchemFP403
PubchemFP404	PubchemFP405	PubchemFP406	PubchemFP407
PubchemFP408	PubchemFP409	PubchemFP411	PubchemFP412
PubchemFP413	PubchemFP414	PubchemFP416	PubchemFP417
PubchemFP418	PubchemFP419	PubchemFP420	PubchemFP421
PubchemFP422	PubchemFP423	PubchemFP425	PubchemFP427
PubchemFP428	PubchemFP429	PubchemFP430	PubchemFP431
PubchemFP432	PubchemFP434	PubchemFP435	PubchemFP436
PubchemFP437	PubchemFP438	PubchemFP439	PubchemFP440
PubchemFP441	PubchemFP442	PubchemFP443	PubchemFP445
PubchemFP446	PubchemFP447	PubchemFP448	PubchemFP449

PubchemFP450	PubchemFP451	PubchemFP452	PubchemFP453
PubchemFP454	PubchemFP456	PubchemFP457	PubchemFP458
PubchemFP459	PubchemFP460	PubchemFP461	PubchemFP462
PubchemFP464	PubchemFP465	PubchemFP466	PubchemFP467
PubchemFP469	PubchemFP470	PubchemFP471	PubchemFP472
PubchemFP473	PubchemFP474	PubchemFP475	PubchemFP476
PubchemFP477	PubchemFP480	PubchemFP481	PubchemFP482
PubchemFP483	PubchemFP484	PubchemFP485	PubchemFP486
PubchemFP487	PubchemFP489	PubchemFP490	PubchemFP493
PubchemFP494	PubchemFP495	PubchemFP497	PubchemFP498
PubchemFP499	PubchemFP500	PubchemFP501	PubchemFP504
PubchemFP505	PubchemFP506	PubchemFP507	PubchemFP508
PubchemFP509	PubchemFP514	PubchemFP515	PubchemFP516
PubchemFP517	PubchemFP518	PubchemFP519	PubchemFP520
PubchemFP521	PubchemFP523	PubchemFP524	PubchemFP530
PubchemFP531	PubchemFP533	PubchemFP534	PubchemFP535
PubchemFP536	PubchemFP537	PubchemFP538	PubchemFP539
PubchemFP540	PubchemFP541	PubchemFP543	PubchemFP544
PubchemFP545	PubchemFP547	PubchemFP548	PubchemFP549
PubchemFP550	PubchemFP551	PubchemFP552	PubchemFP553
PubchemFP554	PubchemFP555	PubchemFP556	PubchemFP558
PubchemFP560	PubchemFP563	PubchemFP564	PubchemFP565
PubchemFP566	PubchemFP567	PubchemFP568	PubchemFP569
PubchemFP570	PubchemFP572	PubchemFP573	PubchemFP574
PubchemFP575	PubchemFP577	PubchemFP578	PubchemFP579
PubchemFP580	PubchemFP581	PubchemFP582	PubchemFP583
PubchemFP584	PubchemFP585	PubchemFP586	PubchemFP588
PubchemFP589	PubchemFP591	PubchemFP592	PubchemFP593
PubchemFP594	PubchemFP595	PubchemFP596	PubchemFP597
PubchemFP598	PubchemFP599	PubchemFP600	PubchemFP601
PubchemFP602	PubchemFP603	PubchemFP604	PubchemFP605
PubchemFP606	PubchemFP607	PubchemFP608	PubchemFP609
PubchemFP610	PubchemFP611	PubchemFP612	PubchemFP613
PubchemFP614	PubchemFP615	PubchemFP616	PubchemFP618
PubchemFP619	PubchemFP620	PubchemFP621	PubchemFP622
PubchemFP623	PubchemFP624	PubchemFP625	PubchemFP626
PubchemFP628	PubchemFP629	PubchemFP630	PubchemFP632
PubchemFP633	PubchemFP634	PubchemFP636	PubchemFP637
PubchemFP638	PubchemFP639	PubchemFP640	PubchemFP641
PubchemFP642	PubchemFP644	PubchemFP645	PubchemFP646
PubchemFP647	PubchemFP648	PubchemFP650	PubchemFP651
PubchemFP652	PubchemFP653	PubchemFP654	PubchemFP655
PubchemFP656	PubchemFP657	PubchemFP658	PubchemFP659
PubchemFP660	PubchemFP661	PubchemFP662	PubchemFP664
PubchemFP665	PubchemFP666	PubchemFP668	PubchemFP669
PubchemFP670	PubchemFP671	PubchemFP673	PubchemFP674
PubchemFP675	PubchemFP676	PubchemFP677	PubchemFP678
PubchemFP679	PubchemFP680	PubchemFP681	PubchemFP682

PubchemFP683	PubchemFP684	PubchemFP685	PubchemFP686
PubchemFP687	PubchemFP688	PubchemFP689	PubchemFP690
PubchemFP691	PubchemFP692	PubchemFP693	PubchemFP694
PubchemFP695	PubchemFP696	PubchemFP697	PubchemFP698
PubchemFP699	PubchemFP700	PubchemFP701	PubchemFP702
PubchemFP703	PubchemFP704	PubchemFP705	PubchemFP706
PubchemFP707	PubchemFP708	PubchemFP709	PubchemFP710
PubchemFP711	PubchemFP712	PubchemFP713	PubchemFP714
PubchemFP715	PubchemFP716	PubchemFP717	PubchemFP719
PubchemFP721	PubchemFP722	PubchemFP725	PubchemFP728
PubchemFP729	PubchemFP733	PubchemFP734	PubchemFP735
PubchemFP736	PubchemFP737	PubchemFP738	PubchemFP740
PubchemFP742	PubchemFP743	PubchemFP745	PubchemFP746
PubchemFP747	PubchemFP749	PubchemFP750	PubchemFP751
PubchemFP752	PubchemFP755	PubchemFP756	PubchemFP757
PubchemFP758	PubchemFP759	PubchemFP761	PubchemFP762
PubchemFP763	PubchemFP764	PubchemFP766	PubchemFP767
PubchemFP770	PubchemFP771	PubchemFP772	PubchemFP776
PubchemFP777	PubchemFP778	PubchemFP779	PubchemFP780
PubchemFP782	PubchemFP784	PubchemFP785	PubchemFP788
PubchemFP791	PubchemFP792	PubchemFP796	PubchemFP797
PubchemFP798	PubchemFP799	PubchemFP800	PubchemFP801
PubchemFP803	PubchemFP805	PubchemFP806	PubchemFP808
PubchemFP809	PubchemFP810	PubchemFP812	PubchemFP813
PubchemFP814	PubchemFP815	PubchemFP818	PubchemFP819
PubchemFP820	PubchemFP821	PubchemFP822	PubchemFP824
PubchemFP825	PubchemFP826	PubchemFP827	PubchemFP829
PubchemFP830	PubchemFP833	PubchemFP834	PubchemFP835
PubchemFP839	PubchemFP840	PubchemFP860	PubchemFP861
nAcid	apol	naAromAtom	nAromBond
nAtom	nHeavyAtom	nH	nB
nC	nN	nO	nS
nP	nF	nCl	nBr
nI	ATSc1	ATSc2	ATSc3
ATSc4	ATSc5	ATSm1	ATSm2
ATSm3	ATSm4	ATSm5	ATSp1
ATSp2	ATSp3	ATSp4	ATSp5
nBase	nBonds	nBonds2	nBondsS
nBondsS2	nBondsS3	nBondsD	nBondsD2
nBondsT	bpol	C1SP1	C2SP1
C1SP2	C2SP2	C3SP2	C1SP3
C2SP3	C3SP3	C4SP3	SCH-3
SCH-4	SCH-5	SCH-6	SCH-7
VCH-3	VCH-4	VCH-5	VCH-6
VCH-7	SC-3	SC-4	SC-5
SC-6	VC-3	VC-4	VC-5
VC-6	SPC-4	SPC-5	SPC-6
VPC-4	VPC-5	VPC-6	ECCEN

fragC	nHBAcc	nHBAcc2	nHBAcc3
nHBAcc_Lipinski	nHBDon	nHBDon_Lipinski	nAtomLC
nAtomP	nAtomLAC	MLogP	McGowan_Volume
MDEC-11	MDEC-12	MDEC-13	MDEC-14
MDEC-22	MDEC-23	MDEC-24	MDEC-33
MDEC-34	MDEC-44	MDEO-11	MDEO-12
MDEO-22	MDEN-11	MDEN-12	MDEN-13
MDEN-22	MDEN-23	MDEN-33	MLFER_A
MLFER_BH	MLFER_BO	MLFER_S	MLFER_E
MLFER_L	PetitjeanNumber	nRing	n3Ring
n4Ring	n5Ring	n6Ring	n7Ring
n8Ring	n9Ring	n10Ring	nG12Ring
nFRing	nF6Ring	nF8Ring	nF9Ring
nF10Ring	nF11Ring	nF12Ring	nFG12Ring
nTRing	nT4Ring	nT5Ring	nT6Ring
nT7Ring	nT8Ring	nT9Ring	nT10Ring
nT11Ring	nT12Ring	nTG12Ring	nRotB
TopoPSA	VAdjMat	MW	WTPT-1
WTPT-2	WTPT-3	WTPT-4	WTPT-5
WPATH	WPOL	Zagreb	

Table A.1: Original Features after preprocessing

Appendix B

Learning Curve Results

	2%					
	Test Fold		Train Set		Test Set	
	RMSE	R^2	RMSE	R^2	RMSE	R^2
1	0,966	0,642	0,466	0,916	1,034	0,584
2	0,833	0,693	0,484	0,91	1,035	0,593
3	0,987	0,589	0,475	0,914	1,037	0,594
4	1,033	0,518	0,457	0,921	1,043	0,586
5	0,929	0,68	0,467	0,915	1,02	0,6
6	0,892	0,721	0,48	0,91	1,03	0,591
7	0,903	0,684	0,48	0,91	1,034	0,589
8	0,87	0,713	0,48	0,911	1,02	0,6
AVG	0,934714286	0,646714286	0,472714286	0,913714286	1,033285714	0,591

Table B.1: Per fold RMSE and R squared results for 2% of the dataset

	20%					
	Test Fold		Train Set		Test Set	
	RMSE	R^2	RMSE	R^2	RMSE	R^2
1	0,829	0,723	0,502	0,896	0,989	0,62
2	0,827	0,711	0,504	0,896	0,997	0,616
3	0,877	0,676	0,491	0,901	0,993	0,618
4	0,858	0,704	0,497	0,898	0,995	0,618
5	0,865	0,685	0,501	0,897	0,922	0,621
6	0,845	0,696	0,503	0,896	0,922	0,619
7	0,86	0,694	0,5	0,897	0,989	0,622
8	0,828	0,706	0,501	0,897	0,992	0,619
AVG	0,851571429	0,698428571	0,499714286	0,897285714	0,972428571	0,619142857

Table B.2: Per fold RMSE and R squared results for 20% of the dataset

	40%					
	Test Fold		Train Set		Test Set	
	RMSE	R^2	RMSE	R^2	RMSE	R^2
1	0,84	0,704	0,534	0,881	0,993	0,614
2	0,842	0,704	0,531	0,881	0,99	0,616
3	0,835	0,712	0,526	0,884	0,992	0,615
4	0,826	0,709	0,532	0,882	0,992	0,614
5	0,862	0,677	0,526	0,884	0,989	0,617
6	0,822	0,723	0,529	0,882	0,99	0,616
7	0,857	0,693	0,529	0,883	0,991	0,615
8	0,848	0,685	0,527	0,884	0,99	0,616
AVG	0,840571429	0,703142857	0,529571429	0,882428571	0,991	0,615285714

Table B.3: Per fold RMSE and R squared results for 40% of the dataset

	60%					
	Test Fold		Train Set		Test Set	
	RMSE	R^2	RMSE	R^2	RMSE	R^2
1	0,82	0,719	0,542	0,878	0,99	0,618
2	0,857	0,703	0,549	0,878	0,992	0,617
3	0,849	0,695	0,54	0,879	0,991	0,618
4	0,832	0,7	0,542	0,878	0,991	0,617
5	0,835	0,713	0,54	0,878	0,986	0,62
6	0,84	0,705	0,54	0,879	0,993	0,616
7	0,816	0,718	0,544	0,877	0,986	0,621
8	0,824	0,709	0,541	0,879	0,986	0,62
AVG	0,835571429	0,707571429	0,542428571	0,878142857	0,989857143	0,618142857

Table B.4: Per fold RMSE and R squared results for 60% of the dataset

	80%					
	Test Fold		Train Set		Test Set	
	RMSE	R^2	RMSE	R^2	RMSE	R^2
1	0,84	0,704	0,554	0,872	0,987	0,618
2	0,841	0,704	0,55	0,874	0,987	0,618
3	0,834	0,712	0,553	0,872	0,988	0,618
4	0,849	0,7	0,552	0,873	0,99	0,616
5	0,806	0,71	0,557	0,872	0,984	0,621
6	0,833	0,713	0,553	0,872	0,987	0,62
7	0,839	0,704	0,552	0,873	0,989	0,618
8	0,829	0,711	0,553	0,873	0,987	0,619
AVG	0,834571429	0,706714286	0,553	0,872571429	0,987428571	0,618428571

Table B.5: Per fold RMSE and R squared results for 80% of the dataset

Appendix C

Boruta Results

Feature Name	**meanImp**	**medianImp**	**minImp**	**maxImp**	**normHits**	**decision**
MS.HL	0,56	0,6	-1,17	2,36	0	Rejected
AKT2_cn	0,56	0,29	-1	2,3	0	Rejected
ALK_wt	0,86	0,95	-1,42	2,48	0	Rejected
APC_wt	7,64	7,6	5,14	9,59	1	Confirmed
APC_cn	1,03	1,4	-0,93	2,07	0	Rejected
BRAF_wt	16,76	16,9	13,77	19,9	1	Confirmed
BRAF_cn	-0,57	-0,69	-3,64	2,14	0	Rejected
BRCA1_wt	1,75	1,74	0	3,37	0	Rejected
BRCA2_wt	0,91	1,03	-1,14	2,35	0	Rejected
BRCA2_cn	0,55	0,75	-0,93	1,6	0	Rejected
CCND1_cn	1,82	1,84	-0,34	3,56	0,01010101	Rejected
CCND2_cn	2,49	2,43	1,22	4,01	0	Rejected
CCND3_cn	-1,43	-1,46	-3,73	0,17	0	Rejected
CDH1_wt	-0,74	-0,61	-2,83	1,52	0	Rejected
CDH1_cn	1,91	2	0,5	3,18	0	Rejected
CDK4_cn	0,88	0,83	-0,66	2,76	0	Rejected
CDK6_cn	-0,58	-0,7	-2	1,38	0	Rejected
CDKN2A_wt	13,93	14	10,74	16,15	1	Confirmed
CDKN2A_cn	12,24	12,36	9,45	14,2	1	Confirmed
CDKN2C_wt	-0,78	-1,01	-2,48	2,23	0	Rejected
CDKN2C_cn	-0,65	-0,89	-2,23	1,33	0	Rejected
CDKN2a.p14._wt	13,26	13,24	10,6	15,13	1	Confirmed
CTNNB1_wt	-0,12	-0,23	-1,57	1,62	0	Rejected
CTNNB1_cn	-0,02	-0,1	-2,8	1,85	0	Rejected
CYLD_wt	0,72	0,99	-1,01	2,05	0	Rejected
EGFR_wt	-0,72	-0,73	-2,21	1,64	0	Rejected
EGFR_cn	-0,79	-0,74	-2,03	0,29	0	Rejected
EP300_cn	0,47	0,56	-2,2	1,7	0	Rejected
ERBB2_wt	0,7	0,69	-0,63	1,83	0	Rejected
ERBB2_cn	7,11	7,11	4,82	9,34	1	Confirmed
EZH2_wt	-0,51	-0,43	-2,31	0,73	0	Rejected
EZH2_cn	-0,45	-0,45	-2,08	0,92	0	Rejected
FAM123B_wt	3,41	3,48	1,65	5,43	0,535353535	Tentative
FAM123B_cn	1,77	1,79	-0,44	2,71	0	Rejected

FBXW7_wt	0,06	-0,05	-2,3	1,86	0	Rejected
FBXW7_cn	1,11	1,38	-0,6	2,68	0	Rejected
FGFR2_cn	2,78	2,92	0,99	4,49	0,090909091	Rejected
FGFR3_wt	1,75	1,81	0,54	2,6	0	Rejected
FGFR3_cn	1,73	1,72	-0,18	2,91	0	Rejected
FLCN_wt	1,02	0,92	-0,49	2,48	0	Rejected
FLT3_wt	1,12	1,13	-1,46	2,55	0	Rejected
FLT3_cn	0	-0,22	-1,62	1,75	0	Rejected
GNAS_wt	1,47	1,33	0,39	2,74	0	Rejected
GNAS_cn	0,73	0,49	-0,92	2,39	0	Rejected
HRAS_wt	-1,89	-1,92	-4,05	0,43	0	Rejected
IDH1_wt	1,18	1,45	-0,53	2,47	0	Rejected
IDH1_cn	0,87	1,12	-0,8	2,24	0	Rejected
JAK2_wt	1,26	1,27	-0,23	2,37	0	Rejected
JAK2_cn	0,29	0,11	-1,48	2,23	0	Rejected
KDM5C_wt	-0,37	-0,29	-2,04	1,92	0	Rejected
KDM5C_cn	-0,01	0	-1,4	1,38	0	Rejected
KDM6A_wt	-0,17	-0,22	-1,62	1,86	0	Rejected
KDM6A_cn	-1,03	-1,02	-2,58	0,43	0	Rejected
KDR_cn	0,87	0,91	-1,17	2,85	0	Rejected
KIT_cn	0,71	0,59	-0,98	1,92	0	Rejected
KRAS_wt	6,76	6,8	4,7	9,2	1	Confirmed
KRAS_cn	0,2	0,4	-1,14	1,05	0	Rejected
MAP2K4_wt	0,14	-0,08	-1,16	2,07	0	Rejected
MAP2K4_cn	-0,31	-0,26	-2,19	1,8	0	Rejected
MDM2_cn	2,18	2,26	0,83	3,41	0	Rejected
MET_cn	0,95	0,94	-0,52	2,36	0	Rejected
MLH1_wt	1,93	1,75	0,03	4,29	0,01010101	Rejected
MLH1_cn	1,55	1,9	-0,93	2,88	0	Rejected
MLLT3_cn	0,96	1,03	-0,92	2,86	0	Rejected
MSH2_wt	-1,72	-1,94	-3,66	0,6	0	Rejected
MSH2_cn	-1,37	-1,39	-3,06	0,31	0	Rejected
MSH6_wt	-1,38	-1,4	-2,69	0,04	0	Rejected
MSH6_cn	0,71	1,02	-1,97	1,92	0	Rejected
MYC_cn	0,63	0,78	-1,36	1,92	0	Rejected
MYCL1_cn	-0,15	-0,31	-1,92	1,39	0	Rejected
MYCN_cn	7,79	7,79	6,26	10,16	1	Confirmed
NF1_wt	-0,54	-0,41	-2,84	0,89	0	Rejected
NF1_cn	0,37	0,66	-1,17	1,82	0	Rejected
NF2_wt	2,46	2,36	-0,01	4,16	0,02020202	Rejected
NF2_cn	-0,34	-0,2	-2,04	1,42	0	Rejected
NOTCH1_wt	3,85	3,67	0,76	8,05	0,585858586	Tentative
NRAS_wt	4,24	4,27	1,32	6,92	0,747474747	Confirmed
NRAS_cn	1,13	1,21	-0,3	2,76	0	Rejected
PDGFRA_cn	0,96	0,89	-1,64	2,85	0	Rejected
PIK3CA_wt	1,91	1,86	0,75	3,01	0	Rejected
PIK3CA_cn	0,69	0,43	-1	3,36	0	Rejected
PIK3R1_wt	1,41	1,57	-0,06	2,51	0	Rejected

PIK3R1_cn	2,63	2,49	0,74	4,8	0,131313131	Rejected
PTEN_wt	1,94	1,6	0,86	4,03	0	Rejected
PTEN_cn	1,51	1,61	-0,41	2,94	0	Rejected
RB1_wt	5,08	5,1	2,61	7,46	0,898989899	Confirmed
RB1_cn	-0,37	-0,24	-1,7	1,24	0	Rejected
RUNX1_wt	0,63	0,67	-0,54	2,03	0	Rejected
SETD2_wt	-0,41	-0,36	-2,12	2,33	0	Rejected
SMAD4_wt	6,89	6,86	4,05	9,19	1	Confirmed
SMAD4_cn	0,64	0,96	-0,62	2,41	0	Rejected
SMARCA4_wt	0,63	0,54	-0,89	2,48	0	Rejected
SMARCA4_cn	0,51	0,78	-1	1,5	0	Rejected
SMARCB1_cn	1,3	1,45	-0,62	3,32	0	Rejected
SMO_cn	-0,79	-0,58	-2,42	0,71	0	Rejected
SOCS1_cn	0,32	0,69	-1,28	1,53	0	Rejected
STK11_wt	1,09	0,88	-0,86	3,28	0	Rejected
STK11_cn	-0,13	-0,29	-2,06	1,5	0	Rejected
TET2_wt	4,54	4,61	2,45	7,24	0,858585859	Confirmed
TP53_wt	11,9	11,88	9,42	14,36	1	Confirmed
TP53_cn	0,24	0,13	-2,15	2,2	0	Rejected
TSC1_wt	1,84	1,62	0,25	3,52	0	Rejected
TSC1_cn	2,15	2,25	-0,2	3,49	0,01010101	Rejected
VHL_wt	1,06	1,13	-0,23	2,23	0	Rejected
VHL_cn	1,27	1,37	-1,22	2,83	0	Rejected
WT1_cn	1,38	1,4	-0,46	2,52	0	Rejected
PubchemFP0	2,08	1,94	1,2	2,94	0	Rejected
PubchemFP1	1,8	1,87	1,06	2,29	0	Rejected
PubchemFP2	-0,23	-0,7	-1,45	1,26	0	Rejected
PubchemFP6	4,8	4,75	3,59	6,43	0,96969697	Confirmed
PubchemFP9	1,67	1,73	1	2,09	0	Rejected
PubchemFP10	1,65	1,61	1	2,65	0	Rejected
PubchemFP11	2,11	2,07	1,51	3,13	0	Rejected
PubchemFP12	2,16	2,13	1,7	2,68	0	Rejected
PubchemFP13	1,68	1,63	0,96	2,37	0	Rejected
PubchemFP14	2,23	2,17	1,34	2,9	0	Rejected
PubchemFP15	2,29	2,18	1,58	3,13	0	Rejected
PubchemFP16	3,75	3,72	2,62	5,04	0,727272727	Confirmed
PubchemFP17	1,57	1,5	0,77	2,39	0	Rejected
PubchemFP18	1,77	1,77	1,3	2,6	0	Rejected
PubchemFP19	3,83	3,87	1,73	5,41	0,787878788	Confirmed
PubchemFP20	4,82	4,85	3,17	5,93	0,949494949	Confirmed
PubchemFP21	2,92	2,96	1,89	3,81	0,292929293	Tentative
PubchemFP22	2,29	2,29	1,66	3,03	0	Rejected
PubchemFP23	3,32	3,26	1,87	5,68	0,464646465	Tentative
PubchemFP24	3,03	2,93	1,82	4,57	0,343434343	Tentative
PubchemFP25	2,06	1,95	1,1	3,18	0	Rejected
PubchemFP30	3,33	3,42	2,11	4,49	0,484848485	Tentative
PubchemFP33	2,98	3,02	2	3,79	0,03030303	Rejected
PubchemFP34	3,4	3,43	2,01	4,43	0,535353535	Tentative

PubchemFP37	3,51	3,56	1,91	5,21	0,575757576	Tentative
PubchemFP38	3,12	3,2	2,39	3,77	0,01010101	Rejected
PubchemFP43	1,7	1,73	1,09	2,27	0	Rejected
PubchemFP44	1,63	1,52	1,11	2,74	0	Rejected
PubchemFP46	2,95	2,9	2,29	4,07	0,060606061	Rejected
PubchemFP93	1,9	1,92	1,27	2,32	0	Rejected
PubchemFP115	2,2	2,08	1,48	3,22	0	Rejected
PubchemFP116	2,41	2,5	1,65	3,11	0	Rejected
PubchemFP117	1,67	1,73	0,93	2,36	0	Rejected
PubchemFP118	1,89	2,08	0,91	2,52	0	Rejected
PubchemFP129	0,73	0,97	-1,12	1,54	0	Rejected
PubchemFP130	0,58	0,74	-1,49	1,82	0	Rejected
PubchemFP132	0,75	0,97	-0,46	1,79	0	Rejected
PubchemFP143	3,72	3,71	1,71	5,82	0,626262626	Tentative
PubchemFP144	2,57	2,56	1,95	3,37	0	Rejected
PubchemFP145	3,2	3,17	1,85	4,83	0,424242424	Tentative
PubchemFP146	3,81	3,79	2,2	5,82	0,666666667	Tentative
PubchemFP147	1,01	1,18	0,02	1,69	0	Rejected
PubchemFP148	3,21	3,18	2,18	4,78	0,393939394	Tentative
PubchemFP149	3,33	3,37	2,34	4,27	0,454545455	Tentative
PubchemFP150	3,39	3,35	2,11	5,53	0,505050505	Tentative
PubchemFP152	3,83	3,86	2,8	4,78	0,686868687	Confirmed
PubchemFP153	3,24	3,25	2,05	4,35	0,434343434	Tentative
PubchemFP155	3,85	3,87	2,83	5,1	0,717171717	Confirmed
PubchemFP156	3,9	3,95	2,6	4,95	0,757575758	Confirmed
PubchemFP157	4,05	4,04	2,95	5,08	0,818181818	Confirmed
PubchemFP159	2,46	2,47	1,64	3,83	0	Rejected
PubchemFP160	2,35	2,32	1,6	3,64	0	Rejected
PubchemFP164	1,58	1,63	1,01	2,22	0	Rejected
PubchemFP167	1,22	1,29	-0,19	1,79	0	Rejected
PubchemFP178	4,31	4,37	3	5,9	0,858585859	Confirmed
PubchemFP179	2,2	2,23	1,62	2,85	0	Rejected
PubchemFP180	3,56	3,58	2,28	4,58	0,616161616	Tentative
PubchemFP181	3,58	3,57	2,56	4,64	0,636363636	Tentative
PubchemFP182	3,5	3,54	2,05	4,5	0,616161616	Tentative
PubchemFP183	2,42	2,54	1,43	2,81	0	Rejected
PubchemFP184	2,41	2,35	1,77	3,11	0	Rejected
PubchemFP185	2,69	2,58	2,14	3,34	0	Rejected
PubchemFP186	3,8	3,78	2,56	4,8	0,737373737	Confirmed
PubchemFP187	4,41	4,42	3,36	5,61	0,939393939	Confirmed
PubchemFP188	4,72	4,65	3,5	6,15	0,95959596	Confirmed
PubchemFP189	1,66	1,65	1,05	2,55	0	Rejected
PubchemFP190	-1,08	-1,04	-2,39	0,34	0	Rejected
PubchemFP191	3,24	3,24	2,03	4,33	0,424242424	Tentative
PubchemFP192	3,46	3,51	1,63	4,75	0,505050505	Tentative
PubchemFP193	3,58	3,6	2,05	4,8	0,666666667	Tentative
PubchemFP194	2,9	2,86	2,07	4,06	0,03030303	Rejected
PubchemFP195	2,71	2,72	1,84	3,43	0,080808081	Rejected

PubchemFP199	2,72	2,81	1,82	3,58	0	Rejected
PubchemFP200	2,25	2,18	1,7	3,03	0	Rejected
PubchemFP206	3,07	3,05	1,86	4,31	0,393939394	Tentative
PubchemFP213	1,85	1,83	1,18	2,63	0	Rejected
PubchemFP214	-0,39	-0,85	-1,84	1,52	0	Rejected
PubchemFP218	1,77	1,82	1,23	2,26	0	Rejected
PubchemFP219	1,75	1,72	1,05	2,29	0	Rejected
PubchemFP227	0,3	0,65	-1,62	1,66	0	Rejected
PubchemFP228	0,96	1	0,13	1,42	0	Rejected
PubchemFP232	-0,35	-0,47	-1,91	1,33	0	Rejected
PubchemFP233	-1,31	-1,64	-2,63	0,42	0	Rejected
PubchemFP241	1,31	1,39	-0,08	2,5	0	Rejected
PubchemFP246	1,36	1,32	-0,44	2,7	0	Rejected
PubchemFP247	1,25	1,29	-0,79	2,41	0	Rejected
PubchemFP248	0,66	0,99	-0,89	1,37	0	Rejected
PubchemFP252	0,96	1,14	-1,03	1,62	0	Rejected
PubchemFP255	2,3	2,27	1,83	3,26	0	Rejected
PubchemFP256	2,79	2,79	1,78	3,82	0,02020202	Rejected
PubchemFP257	3,13	3,08	2,11	4,37	0,373737374	Tentative
PubchemFP258	3,47	3,49	1,95	4,86	0,585858586	Tentative
PubchemFP259	3,15	3,13	1,98	4,21	0,373737374	Tentative
PubchemFP260	2,61	2,56	2,1	3,07	0	Rejected
PubchemFP261	3,6	3,61	2,62	4,6	0,616161616	Tentative
PubchemFP262	1,45	1,65	-0,07	2,53	0	Rejected
PubchemFP274	4,81	4,84	3,28	6,18	0,95959596	Confirmed
PubchemFP276	4,89	4,88	3,1	6,13	0,949494949	Confirmed
PubchemFP283	1,58	1,56	1,04	2,18	0	Rejected
PubchemFP284	1,82	1,78	1,06	2,92	0	Rejected
PubchemFP285	2,82	2,97	1,66	3,32	0	Rejected
PubchemFP286	1,93	1,8	1,22	2,71	0	Rejected
PubchemFP287	3,39	3,35	1,84	5,47	0,505050505	Tentative
PubchemFP293	3,08	3,1	1,74	4,39	0,373737374	Tentative
PubchemFP294	3,06	2,95	1,83	4,52	0,373737374	Tentative
PubchemFP297	1,45	1,35	1	2,06	0	Rejected
PubchemFP298	2,76	2,68	2,26	3,59	0	Rejected
PubchemFP299	2,8	2,92	1,86	3,37	0	Rejected
PubchemFP300	2,5	2,36	1,59	3,95	0,01010101	Rejected
PubchemFP301	4,19	4,2	3,07	5,45	0,858585859	Confirmed
PubchemFP305	1,84	1,85	0,38	2,96	0	Rejected
PubchemFP308	6,43	6,42	4,85	7,53	0,98989899	Confirmed
PubchemFP314	3,23	3,2	2,31	4,34	0,383838384	Tentative
PubchemFP327	1,73	1,64	1,11	2,42	0	Rejected
PubchemFP328	1,72	1,76	1,21	2,63	0	Rejected
PubchemFP330	1,68	1,61	1,28	2,24	0	Rejected
PubchemFP332	2,06	2,09	1,27	2,92	0	Rejected
PubchemFP333	2,33	2,28	1,86	3,4	0	Rejected
PubchemFP334	3,43	3,49	2,3	4,71	0,505050505	Tentative
PubchemFP335	3,63	3,6	2,32	4,62	0,656565657	Tentative

PubchemFP336	2,68	2,71	1,31	3,4	0,02020202	Rejected
PubchemFP337	4,27	4,27	3,31	5,58	0,898989899	Confirmed
PubchemFP338	2,86	2,81	1,81	4,46	0,303030303	Tentative
PubchemFP339	3,27	3,26	2,5	4,38	0,414141414	Tentative
PubchemFP340	2,24	2,13	1,12	3,57	0	Rejected
PubchemFP341	2,68	2,57	1,88	4,24	0,01010101	Rejected
PubchemFP342	2,78	2,84	1,8	3,32	0	Rejected
PubchemFP344	1,71	1,71	1,07	2,46	0	Rejected
PubchemFP345	3,33	3,39	1,92	4,64	0,515151515	Tentative
PubchemFP346	3,63	3,65	2,72	5,24	0,646464646	Tentative
PubchemFP347	1,74	1,69	1,02	2,44	0	Rejected
PubchemFP349	3,02	3,03	1,58	4,46	0,292929293	Tentative
PubchemFP350	2,84	2,96	1,92	3,58	0	Rejected
PubchemFP351	2,19	2,18	1,15	3,26	0	Rejected
PubchemFP352	1,91	1,86	1,23	2,68	0	Rejected
PubchemFP353	3,06	3,04	1,94	4,21	0,343434343	Tentative
PubchemFP355	2,28	2,26	1,5	3,04	0	Rejected
PubchemFP356	2,71	2,79	1,79	4,03	0	Rejected
PubchemFP357	3,51	3,53	2,14	5,26	0,565656566	Tentative
PubchemFP358	2,62	2,61	2,02	3,39	0	Rejected
PubchemFP359	2,64	2,71	1,89	3,37	0,02020202	Rejected
PubchemFP360	-0,27	-0,42	-1,95	1,03	0	Rejected
PubchemFP362	2,89	2,89	2,17	3,74	0,050505051	Rejected
PubchemFP363	2,43	2,47	1,46	3,93	0,01010101	Rejected
PubchemFP364	2,85	2,89	1,86	4,3	0,02020202	Rejected
PubchemFP365	3,3	3,3	1,68	4,48	0,444444444	Tentative
PubchemFP366	4,18	4,24	3,19	5,08	0,868686869	Confirmed
PubchemFP367	1,9	1,91	1,27	2,69	0	Rejected
PubchemFP368	3,35	3,35	2,29	4,47	0,494949495	Tentative
PubchemFP370	2,2	2,14	1,37	3,08	0	Rejected
PubchemFP371	2,03	2,09	1,49	2,6	0	Rejected
PubchemFP372	3,42	3,43	1,88	5,06	0,525252525	Tentative
PubchemFP373	3,18	3,16	2,05	4,74	0,353535354	Tentative
PubchemFP374	3,28	3,28	1,92	4,71	0,454545455	Tentative
PubchemFP375	3,46	3,49	2,21	4,99	0,515151515	Tentative
PubchemFP376	3,18	3,27	2,03	4,16	0,383838384	Tentative
PubchemFP377	3,95	3,99	2,66	5,38	0,747474747	Confirmed
PubchemFP378	2,93	2,8	1,85	4,37	0,282828283	Rejected
PubchemFP379	2,51	2,65	1,87	3,24	0	Rejected
PubchemFP380	3,14	3,18	2,15	3,92	0,383838384	Tentative
PubchemFP381	2,9	2,85	2,43	3,93	0	Rejected
PubchemFP382	2,78	2,73	1,85	3,85	0,01010101	Rejected
PubchemFP383	3,08	3,07	2,01	4,18	0,343434343	Tentative
PubchemFP384	2,12	2,23	1,49	2,69	0	Rejected
PubchemFP385	3,32	3,31	1,93	5,05	0,515151515	Tentative
PubchemFP386	3,63	3,66	1,8	5,37	0,636363636	Tentative
PubchemFP387	2,99	2,99	1,81	4,01	0,313131313	Tentative
PubchemFP388	2,26	2,33	1,2	3,24	0	Rejected

PubchemFP389	3,11	3,13	1,74	4,27	0,393939394	Tentative
PubchemFP390	2,76	2,77	1,78	3,57	0,080808081	Rejected
PubchemFP391	4,07	4,06	2,58	5,08	0,828282828	Confirmed
PubchemFP392	3,41	3,44	2,31	4,64	0,484848485	Tentative
PubchemFP393	3,44	3,43	1,86	4,76	0,545454545	Tentative
PubchemFP394	2,71	2,75	1,89	3,52	0,01010101	Rejected
PubchemFP395	3,97	3,98	2,89	5,07	0,797979798	Confirmed
PubchemFP396	2,82	2,85	1,9	3,89	0,03030303	Rejected
PubchemFP397	2,91	2,94	2,27	3,29	0	Rejected
PubchemFP398	2,66	2,68	1,48	3,7	0,04040404	Rejected
PubchemFP399	2,86	2,88	1,79	3,83	0,090909091	Rejected
PubchemFP400	2,83	2,83	1,82	3,48	0,02020202	Rejected
PubchemFP403	3,67	3,64	2,32	4,89	0,686868687	Tentative
PubchemFP404	1,52	1,55	0,55	2,26	0	Rejected
PubchemFP405	2,41	2,44	1,68	3,28	0	Rejected
PubchemFP406	4,55	4,51	3,64	5,78	0,95959596	Confirmed
PubchemFP407	3,21	3,26	2,1	4,29	0,424242424	Tentative
PubchemFP408	2,99	3	2,18	4,1	0,363636364	Tentative
PubchemFP409	0,46	0,4	-1,18	1,67	0	Rejected
PubchemFP411	3,3	3,33	2,06	4,2	0,434343434	Tentative
PubchemFP412	3,26	3,26	1,97	4,42	0,515151515	Tentative
PubchemFP413	1,62	1,72	0,98	2,03	0	Rejected
PubchemFP414	2,9	2,9	1,75	4,9	0,282828283	Rejected
PubchemFP416	2,32	2,25	1,43	3,13	0	Rejected
PubchemFP417	2,6	2,61	1,71	3,54	0	Rejected
PubchemFP418	3,7	3,73	2,45	5,18	0,676767677	Tentative
PubchemFP419	2,89	2,94	1,57	3,95	0,323232323	Tentative
PubchemFP420	2,58	2,42	1,69	3,56	0	Rejected
PubchemFP421	2,13	2,16	1,2	2,84	0	Rejected
PubchemFP422	3,13	3,14	1,9	3,98	0,373737374	Tentative
PubchemFP423	0,79	1,18	-1,14	1,68	0	Rejected
PubchemFP425	3,23	3,26	1,86	4,22	0,424242424	Tentative
PubchemFP427	2,26	2,14	1,55	3,06	0	Rejected
PubchemFP428	0,96	1,28	-1,72	1,8	0	Rejected
PubchemFP429	2,61	2,43	1,59	3,35	0	Rejected
PubchemFP430	3,22	3,29	2,24	4,29	0,454545455	Tentative
PubchemFP431	3,4	3,43	1,97	4,84	0,525252525	Tentative
PubchemFP432	3,57	3,59	2,13	4,81	0,585858586	Tentative
PubchemFP434	3,29	3,3	1,9	4,39	0,434343434	Tentative
PubchemFP435	3,57	3,54	2,43	5,25	0,626262626	Tentative
PubchemFP436	0,78	0,87	-1,41	1,66	0	Rejected
PubchemFP437	2,81	2,89	1,5	3,24	0	Rejected
PubchemFP438	3,23	3,22	1,45	4,62	0,484848485	Tentative
PubchemFP439	4,58	4,58	2,64	6	0,898989899	Confirmed
PubchemFP440	3,15	3,17	2,3	4,21	0,353535354	Tentative
PubchemFP441	2,3	2,26	1,65	3,09	0	Rejected
PubchemFP442	2,66	2,67	2,01	3,52	0	Rejected
PubchemFP443	2,87	2,9	1,95	4,01	0,080808081	Rejected

PubchemFP445	3,15	3,15	1,68	4,73	0,373737374	Tentative
PubchemFP446	3,57	3,62	2,19	4,87	0,616161616	Tentative
PubchemFP447	3,46	3,47	2	4,65	0,545454545	Tentative
PubchemFP448	0,5	0,76	-1,28	1,64	0	Rejected
PubchemFP449	2,3	2,22	1,47	3,48	0	Rejected
PubchemFP450	3,25	3,24	1,92	5,1	0,444444444	Tentative
PubchemFP451	2,81	2,77	1,79	3,56	0	Rejected
PubchemFP452	2,93	3,06	2,18	3,56	0,02020202	Rejected
PubchemFP453	3,69	3,7	2,24	4,85	0,666666667	Tentative
PubchemFP454	0,93	1	-0,45	1,88	0	Rejected
PubchemFP456	2,96	2,99	2,21	3,64	0	Rejected
PubchemFP457	2,9	2,83	1,92	4,02	0,050505051	Rejected
PubchemFP458	2,93	2,93	2	4	0,101010101	Rejected
PubchemFP459	2,74	2,77	1,61	3,56	0,03030303	Rejected
PubchemFP460	2,29	2,22	1,69	3,23	0	Rejected
PubchemFP461	2,79	2,78	2,09	3,21	0	Rejected
PubchemFP462	3,42	3,48	2,17	4,35	0,555555556	Tentative
PubchemFP464	2,42	2,26	1,37	3,7	0	Rejected
PubchemFP465	3,42	3,42	2,24	4,54	0,494949495	Tentative
PubchemFP466	2,98	3,01	1,91	4,41	0,131313131	Rejected
PubchemFP467	2,9	2,84	1,29	4,7	0,313131313	Tentative
PubchemFP469	0,2	0,75	-1,96	1,64	0	Rejected
PubchemFP470	2,34	2,39	1,09	3,35	0	Rejected
PubchemFP471	2,79	2,9	1,57	3,91	0,060606061	Rejected
PubchemFP472	3,72	3,76	1,99	4,95	0,656565657	Tentative
PubchemFP473	1,1	1,21	-0,28	1,8	0	Rejected
PubchemFP474	0,92	1,07	-1,4	2,08	0	Rejected
PubchemFP475	0,77	1,11	-1,74	2,07	0	Rejected
PubchemFP476	2,64	2,52	1,3	3,64	0,02020202	Rejected
PubchemFP477	3,03	3,06	1,64	4,19	0,383838384	Tentative
PubchemFP480	2,2	2,29	1,02	2,98	0	Rejected
PubchemFP481	1,93	1,95	1,14	2,94	0	Rejected
PubchemFP482	2,63	2,71	1,89	3,33	0,01010101	Rejected
PubchemFP483	1,4	1,25	0,55	2,32	0	Rejected
PubchemFP484	2,82	2,79	1,68	4,12	0,202020202	Rejected
PubchemFP485	3,16	3,22	1,82	4,51	0,424242424	Tentative
PubchemFP486	2,58	2,58	1,8	3,58	0,03030303	Rejected
PubchemFP487	3,19	3,09	2,09	5,26	0,464646465	Tentative
PubchemFP489	1,58	1,62	0,1	2,47	0	Rejected
PubchemFP490	2,27	2,24	1,27	3,08	0	Rejected
PubchemFP493	3,93	3,99	1,86	5,74	0,686868687	Tentative
PubchemFP494	2,9	2,89	2,28	3,62	0,03030303	Rejected
PubchemFP495	2,81	2,82	2,14	3,4	0	Rejected
PubchemFP497	2,86	2,85	2	3,54	0	Rejected
PubchemFP498	3,67	3,71	2,16	5,15	0,676767677	Tentative
PubchemFP499	3,48	3,57	1,8	4,69	0,575757576	Tentative
PubchemFP500	3,05	3,06	1,93	4,24	0,343434343	Tentative
PubchemFP501	2,99	2,9	2,42	3,71	0	Rejected

PubchemFP504	3,06	3,01	1,7	4,27	0,313131313	Tentative
PubchemFP505	1,19	1,32	-0,41	2,07	0	Rejected
PubchemFP506	3,6	3,62	2,22	5,06	0,595959596	Tentative
PubchemFP507	2,97	2,93	2,24	4,11	0	Rejected
PubchemFP508	1,32	1,43	0,12	2,35	0	Rejected
PubchemFP509	1,59	1,52	1	2,21	0	Rejected
PubchemFP514	5,17	5,21	3,58	6,36	0,95959596	Confirmed
PubchemFP515	2,69	2,5	1,96	3,66	0,01010101	Rejected
PubchemFP516	2,18	2,19	1,45	2,83	0	Rejected
PubchemFP517	2,45	2,55	1,17	3,27	0,01010101	Rejected
PubchemFP518	1,57	1,48	0,92	2,42	0	Rejected
PubchemFP519	3,59	3,56	2,18	4,83	0,686868687	Tentative
PubchemFP520	2,23	2,22	1,4	2,91	0	Rejected
PubchemFP521	2,95	2,9	2,19	3,94	0,090909091	Rejected
PubchemFP523	2,62	2,68	1,73	3,43	0	Rejected
PubchemFP524	2,42	2,4	1,64	2,95	0	Rejected
PubchemFP530	2,42	2,57	1,41	3,28	0	Rejected
PubchemFP531	3,22	3,27	1,94	4,41	0,484848485	Tentative
PubchemFP533	1,75	1,92	0,53	2,72	0	Rejected
PubchemFP534	2,44	2,47	1,26	3,07	0	Rejected
PubchemFP535	2,94	2,92	1,92	3,95	0,03030303	Rejected
PubchemFP536	4,06	3,99	2,98	5,21	0,828282828	Confirmed
PubchemFP537	3,45	3,42	2,38	4,5	0,535353535	Tentative
PubchemFP538	2,65	2,59	1,86	3,48	0	Rejected
PubchemFP539	-0,8	-1,06	-1,7	1,23	0	Rejected
PubchemFP540	3,32	3,32	2,25	4,58	0,494949495	Tentative
PubchemFP541	3,11	3,09	1,82	4,59	0,323232323	Tentative
PubchemFP543	2,76	2,61	2,04	3,42	0	Rejected
PubchemFP544	2,62	2,66	1,48	3,66	0,01010101	Rejected
PubchemFP545	2,93	2,9	2,15	3,95	0,03030303	Rejected
PubchemFP547	3,53	3,44	2,25	5,19	0,565656566	Tentative
PubchemFP548	3,13	3,14	1,91	4,74	0,414141414	Tentative
PubchemFP549	2,16	2,16	1,53	2,66	0	Rejected
PubchemFP550	3,05	3,03	2,04	4,27	0,333333333	Tentative
PubchemFP551	1,36	1,23	0,82	2,18	0	Rejected
PubchemFP552	2	1,93	1,17	2,88	0	Rejected
PubchemFP553	3,67	3,65	2,27	5,06	0,626262626	Tentative
PubchemFP554	1,53	1,45	1	2,34	0	Rejected
PubchemFP555	2,91	3,05	2,02	3,74	0,02020202	Rejected
PubchemFP556	2,5	2,42	1,92	3,42	0	Rejected
PubchemFP558	0,43	0,77	-1,18	1,8	0	Rejected
PubchemFP560	3,25	3,13	1,82	5,04	0,363636364	Tentative
PubchemFP563	2,57	2,6	1,91	3,24	0	Rejected
PubchemFP564	1,99	1,98	1,22	2,54	0	Rejected
PubchemFP565	2,42	2,47	1,76	3,16	0	Rejected
PubchemFP566	3,65	3,6	2,67	4,82	0,616161616	Tentative
PubchemFP567	3,21	3,18	2,09	4,13	0,414141414	Tentative
PubchemFP568	2,66	2,7	2,13	3,17	0	Rejected

PubchemFP569	3,29	3,36	1,98	4,34	0,484848485	Tentative
PubchemFP570	2,03	2,01	1,11	2,48	0	Rejected
PubchemFP572	2,77	2,77	1,81	4,03	0,090909091	Rejected
PubchemFP573	2,99	3,07	2,17	3,78	0,01010101	Rejected
PubchemFP574	3,16	3,15	1,87	4,23	0,434343434	Tentative
PubchemFP575	2,25	2,22	1,42	3,17	0	Rejected
PubchemFP577	3,37	3,43	1,89	4,63	0,545454545	Tentative
PubchemFP578	2,05	2,05	1,55	2,62	0	Rejected
PubchemFP579	2,52	2,48	1,8	3,46	0	Rejected
PubchemFP580	2,75	2,66	1,68	3,75	0,02020202	Rejected
PubchemFP581	6,53	6,58	4,64	7,7	1	Confirmed
PubchemFP582	2,17	2,22	1,57	2,61	0	Rejected
PubchemFP583	1,37	1,59	-0,03	1,92	0	Rejected
PubchemFP584	2,11	2,06	1,62	2,95	0	Rejected
PubchemFP585	3,08	3,13	1,98	4,47	0,373737374	Tentative
PubchemFP586	3,35	3,36	2,28	4,48	0,444444444	Tentative
PubchemFP588	0,61	0,96	-1,33	1,39	0	Rejected
PubchemFP589	2,51	2,69	1,37	2,98	0	Rejected
PubchemFP591	3,01	2,98	1,71	4,44	0,363636364	Tentative
PubchemFP592	2,01	1,89	1,12	3,4	0,01010101	Rejected
PubchemFP593	3,75	3,76	1,53	5,01	0,686868687	Tentative
PubchemFP594	2,7	2,62	2,11	3,56	0,01010101	Rejected
PubchemFP595	2,13	2,12	1,3	2,85	0	Rejected
PubchemFP596	2,76	2,68	2,18	3,44	0,01010101	Rejected
PubchemFP597	3,33	3,21	1,52	5,4	0,373737374	Tentative
PubchemFP598	2,81	2,76	2,16	3,33	0	Rejected
PubchemFP599	2,16	2	1,46	3,1	0	Rejected
PubchemFP600	3,45	3,52	2,03	4,58	0,555555556	Tentative
PubchemFP601	3,32	3,34	1,97	4,49	0,525252525	Tentative
PubchemFP602	4	4,03	2,07	5,26	0,777777778	Confirmed
PubchemFP603	2,13	2,12	1,43	2,71	0	Rejected
PubchemFP604	2,91	2,9	1,66	4,38	0,303030303	Tentative
PubchemFP605	1,94	2,01	1,12	2,79	0	Rejected
PubchemFP606	2,88	2,93	2,15	3,58	0	Rejected
PubchemFP607	2,25	2,23	1,34	3,18	0	Rejected
PubchemFP608	2,17	2,16	1,33	2,74	0	Rejected
PubchemFP609	1,02	1,02	-0,26	2,23	0	Rejected
PubchemFP610	1,86	1,79	1,07	2,59	0	Rejected
PubchemFP611	3,26	3,27	2,07	4,1	0,474747475	Tentative
PubchemFP612	2,84	2,9	1,55	3,63	0,01010101	Rejected
PubchemFP613	1,96	1,93	1,21	2,7	0	Rejected
PubchemFP614	2,71	2,64	2,22	3,31	0	Rejected
PubchemFP615	3,58	3,55	2,63	4,52	0,616161616	Tentative
PubchemFP616	2,97	2,96	1,95	3,86	0,161616162	Rejected
PubchemFP618	2,05	2,05	1,65	2,67	0	Rejected
PubchemFP619	3,25	3,25	1,92	4,87	0,454545455	Tentative
PubchemFP620	2,93	2,89	2,2	3,98	0,101010101	Rejected
PubchemFP621	3,26	3,23	2,07	4,68	0,424242424	Tentative

PubchemFP622	-1,45	-1,69	-2,39	0,62	0	Rejected
PubchemFP623	3,86	3,94	2,13	5,1	0,757575758	Confirmed
PubchemFP624	2,42	2,34	1,43	3,53	0	Rejected
PubchemFP625	2,64	2,65	1,8	3,26	0	Rejected
PubchemFP626	2,53	2,47	2,03	3,3	0	Rejected
PubchemFP628	2,26	2,31	1,69	2,91	0	Rejected
PubchemFP629	1,4	1,54	-0,35	2,4	0	Rejected
PubchemFP630	2,28	2,2	1,73	3,16	0	Rejected
PubchemFP632	2,82	2,78	1,66	3,73	0,01010101	Rejected
PubchemFP633	2,32	2,3	1,97	2,95	0	Rejected
PubchemFP634	2,05	1,97	1,47	2,6	0	Rejected
PubchemFP636	3,13	3,14	2,08	3,97	0,404040404	Tentative
PubchemFP637	3,06	3,12	1,37	4,46	0,414141414	Tentative
PubchemFP638	2,57	2,52	1,93	3,44	0	Rejected
PubchemFP639	3,2	3,17	1,97	4,12	0,404040404	Tentative
PubchemFP640	2,19	2,11	1,22	3,03	0	Rejected
PubchemFP641	3,42	3,4	2,07	4,67	0,545454545	Tentative
PubchemFP642	6,68	6,66	5,12	8,11	0,98989899	Confirmed
PubchemFP644	2,39	2,22	1,58	3,44	0,01010101	Rejected
PubchemFP645	3,08	2,9	1,49	5,24	0,292929293	Tentative
PubchemFP646	3,77	3,84	2,66	4,99	0,696969697	Confirmed
PubchemFP647	2,31	2,21	1,4	3,36	0	Rejected
PubchemFP648	0,07	0,61	-1,72	1,46	0	Rejected
PubchemFP650	2,11	2,08	1,55	2,78	0	Rejected
PubchemFP651	2,89	2,93	1,77	4,08	0,101010101	Rejected
PubchemFP652	2,57	2,63	1,94	3,16	0	Rejected
PubchemFP653	2,07	2,17	1,08	2,91	0	Rejected
PubchemFP654	2,59	2,55	1,89	3,22	0	Rejected
PubchemFP655	3,51	3,54	1,77	4,62	0,575757576	Tentative
PubchemFP656	2,17	2,11	1,34	3,36	0	Rejected
PubchemFP657	2,96	3	2,17	3,54	0,01010101	Rejected
PubchemFP658	3,2	3,25	1,64	4,3	0,444444444	Tentative
PubchemFP659	3,59	3,54	2,15	5,1	0,616161616	Tentative
PubchemFP660	2,39	2,38	1,48	3,2	0,02020202	Rejected
PubchemFP661	3,08	3,09	1,79	4,1	0,373737374	Tentative
PubchemFP662	2,75	2,7	1,71	3,99	0,101010101	Rejected
PubchemFP664	2,02	1,95	1,12	3,16	0	Rejected
PubchemFP665	3,4	3,43	2,31	4,66	0,505050505	Tentative
PubchemFP666	2,81	2,77	1,75	3,75	0,090909091	Rejected
PubchemFP668	1,93	1,93	1,33	2,46	0	Rejected
PubchemFP669	2,16	2,17	1,19	2,93	0	Rejected
PubchemFP670	1,61	1,61	1,11	2,76	0	Rejected
PubchemFP671	3,55	3,53	2,25	4,66	0,626262626	Tentative
PubchemFP673	3,29	3,28	2,24	4,77	0,404040404	Tentative
PubchemFP674	2,76	2,76	1,96	3,55	0	Rejected
PubchemFP675	1,23	1,3	-1,29	2,09	0	Rejected
PubchemFP676	2,61	2,73	1,88	3,32	0	Rejected
PubchemFP677	2,36	2,4	1,61	3,14	0	Rejected

PubchemFP678	2,1	2,01	1,34	3,27	0	Rejected
PubchemFP679	1,84	1,74	1,23	2,4	0	Rejected
PubchemFP680	2,83	2,56	2,04	4,37	0	Rejected
PubchemFP681	3,07	3,05	1,91	4,33	0,333333333	Tentative
PubchemFP682	3,6	3,7	2,34	4,82	0,646464646	Tentative
PubchemFP683	2,02	2,06	1,29	3,29	0,01010101	Rejected
PubchemFP684	3,12	2,93	1,63	5	0,373737374	Tentative
PubchemFP685	3,16	3,18	1,75	4,41	0,434343434	Tentative
PubchemFP686	3,78	3,76	2,62	5,19	0,656565657	Tentative
PubchemFP687	3,87	3,88	2,55	5,21	0,767676768	Confirmed
PubchemFP688	4,94	4,97	3,69	6,18	0,96969697	Confirmed
PubchemFP689	2,48	2,44	1,56	3,77	0,01010101	Rejected
PubchemFP690	2,87	2,84	1,6	4,11	0,121212121	Rejected
PubchemFP691	3,4	3,42	1,88	4,85	0,515151515	Tentative
PubchemFP692	3,42	3,46	1,69	5,78	0,515151515	Tentative
PubchemFP693	2,47	2,38	1,72	3,69	0	Rejected
PubchemFP694	2,13	2,06	1,36	3,46	0	Rejected
PubchemFP695	3,66	3,65	2,55	5,42	0,676767677	Tentative
PubchemFP696	3,67	3,68	2,5	4,82	0,676767677	Tentative
PubchemFP697	3,23	3,17	1,56	5,58	0,414141414	Tentative
PubchemFP698	4,04	4,08	2,53	5,45	0,787878788	Confirmed
PubchemFP699	2,69	2,69	1,57	3,74	0	Rejected
PubchemFP700	2,15	2,24	1,49	2,62	0	Rejected
PubchemFP701	2,57	2,45	1,43	3,67	0,04040404	Rejected
PubchemFP702	3,27	3,24	2,03	4,93	0,434343434	Tentative
PubchemFP703	3,46	3,45	2,03	4,85	0,525252525	Tentative
PubchemFP704	3,19	2,98	1,54	5,51	0,414141414	Tentative
PubchemFP705	2,47	2,47	1,61	3,33	0	Rejected
PubchemFP706	2,24	2,12	1,53	3	0	Rejected
PubchemFP707	3,27	3,27	2,12	4,54	0,454545455	Tentative
PubchemFP708	2,34	2,31	1,6	2,98	0	Rejected
PubchemFP709	2,89	2,95	1,78	4,3	0,090909091	Rejected
PubchemFP710	3,22	3,2	1,93	4,52	0,424242424	Tentative
PubchemFP711	3,45	3,43	2,35	4,6	0,545454545	Tentative
PubchemFP712	2,97	3,02	2,26	3,73	0	Rejected
PubchemFP713	3,76	3,76	2,74	5,36	0,676767677	Tentative
PubchemFP714	3,19	3,21	1,87	5,06	0,424242424	Tentative
PubchemFP715	2,46	2,56	1,74	3,37	0	Rejected
PubchemFP716	4,2	4,18	2,96	5,47	0,868686869	Confirmed
PubchemFP717	2,07	2,1	1,34	2,98	0	Rejected
PubchemFP719	2,85	2,81	1,98	3,88	0,101010101	Rejected
PubchemFP721	2,96	2,94	1,52	4,54	0,292929293	Tentative
PubchemFP722	1,53	1,55	0,79	2,28	0	Rejected
PubchemFP725	1,55	1,59	0,87	2,17	0	Rejected
PubchemFP728	1,96	1,97	1,31	2,67	0	Rejected
PubchemFP729	2,28	2,06	1,68	3,04	0	Rejected
PubchemFP733	1,53	1,48	1	2,04	0	Rejected
PubchemFP734	4	3,95	2,99	5,24	0,808080808	Confirmed

PubchemFP735	3,67	3,57	2,11	5,16	0,626262626	Tentative
PubchemFP736	3,41	3,46	2,13	4,82	0,525252525	Tentative
PubchemFP737	3,15	3,09	2,26	4,3	0,131313131	Rejected
PubchemFP738	0,58	0,49	-0,59	2,12	0	Rejected
PubchemFP740	1,88	1,92	0,87	2,69	0	Rejected
PubchemFP742	3,15	3,21	1,77	4,11	0,404040404	Tentative
PubchemFP743	0,52	0,87	-0,99	1,69	0	Rejected
PubchemFP745	1,33	1,33	0,17	2,05	0	Rejected
PubchemFP746	1,57	1,73	0,38	2,28	0	Rejected
PubchemFP747	1,94	1,96	0,06	2,72	0	Rejected
PubchemFP749	2,94	3,02	1,88	3,77	0,090909091	Rejected
PubchemFP750	3,71	3,74	2,5	5,16	0,686868687	Tentative
PubchemFP751	1,63	1,64	1,03	2,36	0	Rejected
PubchemFP752	2,33	2,25	1,3	3,37	0	Rejected
PubchemFP755	3,51	3,51	2,22	5,14	0,585858586	Tentative
PubchemFP756	3,35	3,26	2,38	4,53	0,444444444	Tentative
PubchemFP757	2,17	2,11	1,67	3,33	0	Rejected
PubchemFP758	3,74	3,71	2,17	4,88	0,666666667	Tentative
PubchemFP759	2,97	2,93	1,48	4,16	0,313131313	Tentative
PubchemFP761	2,82	2,79	1,89	3,82	0,050505051	Rejected
PubchemFP762	0,06	0,41	-2,25	1,78	0	Rejected
PubchemFP763	1,35	1,54	-0,07	2,11	0	Rejected
PubchemFP764	1,15	1,17	0,52	1,93	0	Rejected
PubchemFP766	1,52	1,51	1	2,23	0	Rejected
PubchemFP767	1,77	1,81	0,12	2,35	0	Rejected
PubchemFP770	2,75	2,79	1,96	3,42	0,01010101	Rejected
PubchemFP771	0,82	1,11	-1,17	2,15	0	Rejected
PubchemFP772	1,64	1,58	1,4	2,01	0	Rejected
PubchemFP776	3,81	3,83	2,66	5,14	0,727272727	Confirmed
PubchemFP777	3,05	3,04	1,65	4,2	0,333333333	Tentative
PubchemFP778	2,49	2,6	1,78	3,1	0	Rejected
PubchemFP779	4,12	4,17	2,04	5,35	0,828282828	Confirmed
PubchemFP780	2,2	2,11	1,54	3,25	0	Rejected
PubchemFP782	2,58	2,64	1,95	3,21	0	Rejected
PubchemFP784	3,07	3,07	1,68	4,85	0,343434343	Tentative
PubchemFP785	0,95	1,08	-1,54	2,01	0	Rejected
PubchemFP788	1,57	1,64	0,87	2,11	0	Rejected
PubchemFP791	2,19	2,27	1,23	2,91	0	Rejected
PubchemFP792	1,5	1,37	0,73	2,34	0	Rejected
PubchemFP796	1,51	1,46	1,1	2,32	0	Rejected
PubchemFP797	3,82	3,78	2,08	5,13	0,757575758	Confirmed
PubchemFP798	2,59	2,68	1,38	4,17	0	Rejected
PubchemFP799	3,38	3,36	2,35	4,52	0,494949495	Tentative
PubchemFP800	3,38	3,35	2,56	4,49	0,515151515	Tentative
PubchemFP801	0,28	0,23	-1,01	1,84	0	Rejected
PubchemFP803	1,75	1,58	1,16	2,81	0	Rejected
PubchemFP805	2,94	2,96	1,63	4,03	0,131313131	Rejected
PubchemFP806	1,05	1,12	-0,14	1,75	0	Rejected

PubchemFP808	0,74	1,18	-1,73	1,75	0	Rejected
PubchemFP809	1,79	1,88	1,04	2,49	0	Rejected
PubchemFP810	2,16	2,1	1,67	2,77	0	Rejected
PubchemFP812	3,16	3,16	1,83	4,23	0,383838384	Tentative
PubchemFP813	3,72	3,73	2,43	5,25	0,686868687	Tentative
PubchemFP814	1,4	1,35	1,08	1,79	0	Rejected
PubchemFP815	2,18	2,21	1,53	3,11	0	Rejected
PubchemFP818	3,34	3,35	1,93	4,72	0,535353535	Tentative
PubchemFP819	3,43	3,44	2,35	4,71	0,565656566	Tentative
PubchemFP820	1,96	1,98	1,33	2,42	0	Rejected
PubchemFP821	3,85	3,86	2,03	5,25	0,717171717	Confirmed
PubchemFP822	3,08	3	1,64	4,4	0,373737374	Tentative
PubchemFP824	1,95	1,89	1,28	3,25	0	Rejected
PubchemFP825	0,15	0,46	-1,73	1,26	0	Rejected
PubchemFP826	2,06	1,96	0,76	2,97	0	Rejected
PubchemFP827	1,05	1,02	0,07	2,06	0	Rejected
PubchemFP829	1,61	1,52	1,12	2,76	0	Rejected
PubchemFP830	1,85	2,02	0,25	2,62	0	Rejected
PubchemFP833	2,73	2,81	2,12	3,57	0	Rejected
PubchemFP834	0,92	0,86	-1,32	2,3	0	Rejected
PubchemFP835	1,41	1,36	1,05	1,95	0	Rejected
PubchemFP839	1,77	1,68	1,25	2,7	0	Rejected
PubchemFP840	0,54	0,87	-2,45	1,51	0	Rejected
PubchemFP860	1,58	1,39	0,97	2,4	0	Rejected
PubchemFP861	-0,03	0,25	-1,71	1,12	0	Rejected
nAcid	2,98	2,91	2,33	3,63	0,060606061	Rejected
apol	6,77	6,78	5,32	8,11	1	Confirmed
naAromAtom	6,51	6,48	5,39	7,97	1	Confirmed
nAromBond	6,37	6,47	4,53	7,61	0,98989899	Confirmed
nAtom	6,79	6,82	5,89	7,89	1	Confirmed
nHeavyAtom	5,99	6,01	4,94	7,22	0,98989899	Confirmed
nH	7,17	7,2	6,01	8,55	1	Confirmed
nB	4,81	4,81	3,66	5,96	0,949494949	Confirmed
nC	7,01	6,94	5,52	8,66	0,98989899	Confirmed
nN	5,31	5,29	3,85	6,71	0,96969697	Confirmed
nO	6,94	6,95	5,8	7,9	1	Confirmed
nS	4,16	4,26	2,31	5,81	0,777777778	Confirmed
nP	3,33	3,31	1,56	4,85	0,494949495	Tentative
nF	3,87	3,91	2,2	5,47	0,676767677	Tentative
nCl	4,01	3,97	2,3	5,45	0,777777778	Confirmed
nBr	1,64	1,57	1,23	2,2	0	Rejected
nI	2,76	2,78	2,18	3,38	0	Rejected
ATSc1	7,22	7,23	5,94	8,44	1	Confirmed
ATSc2	7,24	7,28	6,26	8,2	1	Confirmed
ATSc3	9,88	9,89	8,27	11,48	1	Confirmed
ATSc4	8,78	8,81	7,3	10,69	1	Confirmed
ATSc5	9,23	9,15	7,78	10,93	1	Confirmed
ATSm1	7,86	7,89	6,09	9,66	1	Confirmed

ATSm2	6,47	6,47	5,08	8	0,98989899	Confirmed
ATSm3	6,35	6,38	5,24	7,73	0,98989899	Confirmed
ATSm4	6,91	6,91	5,68	8,09	0,98989899	Confirmed
ATSm5	6,63	6,68	5,27	8,04	1	Confirmed
ATSp1	6,74	6,8	4,9	8,3	1	Confirmed
ATSp2	6,73	6,76	5,32	7,93	1	Confirmed
ATSp3	6,66	6,66	5,4	8,15	0,98989899	Confirmed
ATSp4	7,17	7,2	6,23	8,27	1	Confirmed
ATSp5	6,76	6,75	5,56	7,79	1	Confirmed
nBase	5,94	5,92	4,42	7,22	0,98989899	Confirmed
nBonds	6,4	6,38	5,29	7,46	1	Confirmed
nBonds2	6,19	6,21	4,56	7,55	1	Confirmed
nBondsS	6,4	6,39	5,08	7,57	1	Confirmed
nBondsS2	7,44	7,43	6,01	8,59	1	Confirmed
nBondsS3	7,28	7,24	6,29	8,36	1	Confirmed
nBondsD	6,44	6,46	4,33	7,97	0,98989899	Confirmed
nBondsD2	6,05	6,02	4,61	8,06	0,98989899	Confirmed
nBondsT	2,79	2,71	2,06	3,89	0,01010101	Rejected
bpol	6,57	6,53	5,48	7,96	0,98989899	Confirmed
C1SP1	2,73	2,89	1,6	3,47	0	Rejected
C2SP1	2,44	2,55	1,7	3,32	0	Rejected
C1SP2	5,76	5,8	3,77	7,05	1	Confirmed
C2SP2	6,95	6,89	5,13	8,45	0,98989899	Confirmed
C3SP2	7,07	7,05	5,41	8,67	1	Confirmed
C1SP3	5,21	5,05	3,79	7,86	0,97979798	Confirmed
C2SP3	6,12	6,08	5,06	7,3	0,98989899	Confirmed
C3SP3	4,13	4,11	3,1	5,1	0,878787879	Confirmed
C4SP3	3,43	3,51	2,08	4,86	0,525252525	Tentative
SCH.3	2,29	2,18	1,35	4,15	0,01010101	Rejected
SCH.4	3,24	3,25	2,32	4,57	0,454545455	Tentative
SCH.5	5,58	5,63	3,99	6,9	0,96969697	Confirmed
SCH.6	8,02	7,97	6,68	9,72	1	Confirmed
SCH.7	7,27	7,2	5,66	8,66	1	Confirmed
VCH.3	2,44	2,48	1,65	3,47	0	Rejected
VCH.4	2,32	2,31	1,58	3,06	0	Rejected
VCH.5	6,35	6,34	4,5	7,77	1	Confirmed
VCH.6	7,16	7,13	5,17	9,28	1	Confirmed
VCH.7	7,48	7,47	5,69	8,85	1	Confirmed
SC.3	5,84	5,82	4,61	7,21	0,98989899	Confirmed
SC.4	6,18	6,2	4,94	7,3	1	Confirmed
SC.5	6,2	6,23	5,06	7,17	0,98989899	Confirmed
SC.6	4,9	4,87	2,94	6,45	0,939393939	Confirmed
VC.3	5,55	5,49	4,49	7,17	0,98989899	Confirmed
VC.4	6,46	6,46	4,82	8,09	0,98989899	Confirmed
VC.5	6,97	7,05	5,49	8,34	1	Confirmed
VC.6	6,01	5,93	4,63	7,3	0,98989899	Confirmed
SPC.4	5,89	5,93	4,92	6,7	0,98989899	Confirmed
SPC.5	7,41	7,42	6,16	8,83	1	Confirmed

SPC.6	7,03	7,05	6,09	8,08	1	Confirmed
VPC.4	6,46	6,45	5,31	7,58	0,98989899	Confirmed
VPC.5	6,54	6,58	5,02	7,66	0,98989899	Confirmed
VPC.6	6,98	6,94	5,65	8,33	1	Confirmed
ECCEN	6,75	6,78	5,01	8,26	1	Confirmed
fragC	6,76	6,73	5,72	8,03	1	Confirmed
nHBAcc	5,92	5,95	4,91	7,08	1	Confirmed
nHBAcc2	6,15	6,18	4,97	7,22	0,98989899	Confirmed
nHBAcc3	6,5	6,51	4,96	7,82	1	Confirmed
nHBAcc_Lipinski	5,19	5,21	4,11	6,32	0,97979798	Confirmed
nHBDon	8,83	8,95	7,33	10,52	1	Confirmed
nHBDon_Lipinski	7,44	7,5	5,61	8,44	1	Confirmed
nAtomLC	7,39	7,46	6	8,99	1	Confirmed
nAtomP	6,11	6,08	4,99	7,47	0,98989899	Confirmed
nAtomLAC	5,68	5,68	4,12	6,88	0,98989899	Confirmed
MLogP	8,24	8,3	5,83	9,13	1	Confirmed
McGowan_Volume	7,04	7,08	5,57	8,34	1	Confirmed
MDEC.11	6	5,99	4,82	7,33	0,98989899	Confirmed
MDEC.12	6,49	6,52	5,26	7,63	1	Confirmed
MDEC.13	6,3	6,27	5,37	7,46	0,98989899	Confirmed
MDEC.14	5,84	5,83	3,84	7,21	0,98989899	Confirmed
MDEC.22	7,99	8,04	6,55	9,17	1	Confirmed
MDEC.23	6,92	6,99	5,79	7,85	1	Confirmed
MDEC.24	5,24	5,22	3,71	6,44	0,97979798	Confirmed
MDEC.33	6,92	6,94	6,05	7,81	1	Confirmed
MDEC.34	7,45	7,5	5,4	8,83	1	Confirmed
MDEC.44	4,13	4,09	3,21	5,38	0,838383838	Confirmed
MDEO.11	7,62	7,63	6,29	8,82	1	Confirmed
MDEO.12	9,74	9,88	7,4	11,51	1	Confirmed
MDEO.22	4,03	4,08	3,09	5,19	0,818181818	Confirmed
MDEN.11	4,95	4,98	3,45	5,9	0,97979798	Confirmed
MDEN.12	4,27	4,2	2,89	5,85	0,848484848	Confirmed
MDEN.13	6,41	6,49	5,04	7,45	0,98989899	Confirmed
MDEN.22	6,04	6,16	4,33	7,31	0,98989899	Confirmed
MDEN.23	5,21	5,14	3,88	7,25	0,97979798	Confirmed
MDEN.33	4,12	4,13	2,94	5,04	0,828282828	Confirmed
MLFER_A	8,01	8,06	6,73	9,55	1	Confirmed
MLFER_BH	7,29	7,34	5,64	8,64	1	Confirmed
MLFER_BO	7,75	7,76	6,38	8,84	1	Confirmed
MLFER_S	6,97	6,99	5,35	8,99	1	Confirmed
MLFER_E	6,83	6,88	5,55	7,88	1	Confirmed
MLFER_L	8,08	8,21	6	9,28	1	Confirmed
PetitjeanNumber	6,83	6,79	5,33	8,03	1	Confirmed
nRing	4,91	4,93	3,57	6,25	0,97979798	Confirmed
n3Ring	2,17	1,93	1,39	3,16	0	Rejected
n4Ring	1,03	1,19	-1,77	2,08	0	Rejected
n5Ring	5,55	5,57	4,18	6,58	0,98989899	Confirmed
n6Ring	5,59	5,59	4,4	6,73	0,98989899	Confirmed

n7Ring	1,77	1,79	0,95	2,63	0	Rejected
n8Ring	1,05	1,35	-2,02	1,71	0	Rejected
n9Ring	1,38	1,29	0,36	2,11	0	Rejected
n10Ring	0,61	0,99	-1,42	1,88	0	Rejected
nG12Ring	2,64	2,63	1,98	3,19	0	Rejected
nFRing	5,36	5,34	4,11	6,46	0,97979798	Confirmed
nF6Ring	1,62	1,66	-0,2	2,22	0	Rejected
nF8Ring	1,91	1,87	1,41	2,42	0	Rejected
nF9Ring	4,1	4	2,72	6,21	0,747474747	Confirmed
nF10Ring	4,66	4,67	3,19	6,06	0,919191919	Confirmed
nF11Ring	1,79	1,78	1,25	2,47	0	Rejected
nF12Ring	2,76	2,9	1,46	3,32	0	Rejected
nFG12Ring	5,35	5,3	4,34	6,38	0,98989899	Confirmed
nTRing	6,31	6,33	5,13	7,4	0,98989899	Confirmed
nT4Ring	0,7	0,99	-1,32	1,69	0	Rejected
nT5Ring	5,38	5,33	3,76	6,56	0,98989899	Confirmed
nT6Ring	5,55	5,62	4,27	6,43	0,98989899	Confirmed
nT7Ring	1,73	1,76	0,85	2,26	0	Rejected
nT8Ring	2,04	2,11	1,19	3,04	0	Rejected
nT9Ring	4,22	4,17	2,96	6,74	0,828282828	Confirmed
nT10Ring	4,7	4,66	3,16	6,35	0,949494949	Confirmed
nT11Ring	1,93	1,91	1,26	2,9	0	Rejected
nT12Ring	2,68	2,66	1,74	4,24	0,03030303	Rejected
nTG12Ring	5,58	5,59	4,58	6,58	0,98989899	Confirmed
nRotB	6,21	6,22	4,66	7,58	0,98989899	Confirmed
TopoPSA	6,56	6,48	5,03	8,32	0,98989899	Confirmed
VAdjMat	5,98	6,01	4,68	7,26	1	Confirmed
MW	7,73	7,71	6,14	9,8	1	Confirmed
WTPT.1	6,93	6,97	5,24	8,14	1	Confirmed
WTPT.2	6,84	6,86	5,88	7,83	0,98989899	Confirmed
WTPT.3	6,3	6,28	5,19	7,53	0,98989899	Confirmed
WTPT.4	8,28	8,28	6,49	9,44	1	Confirmed
WTPT.5	6,83	6,76	5,68	8,13	1	Confirmed
WPATH	7,95	7,9	6,37	9,8	1	Confirmed
WPOL	6,63	6,65	5,42	8,03	0,98989899	Confirmed
Zagreb	6,42	6,39	4,73	7,74	0,98989899	Confirmed

Table C.1: Full Boruta results with importance and reject decision

Appendix D

Feature Importance for gSpan folds

Feature	IncMSE	Feature	IncMSE	Feature	IncMSE
MS.HL	56,33551731	AKT2_cn	16,05456617	ALK_wt	-1,855406336
APC_wt	60,56943559	APC_cn	9,512828573	BRAF_wt	121,8226432
BRAF_cn	-0,958455628	BRCA1_wt	22,68746517	BRCA2_wt	8,978954513
BRCA2_cn	-0,339519667	CCND1_cn	29,05060436	CCND2_cn	29,49272811
CCND3_cn	0,267537424	CDH1_wt	-16,14957804	CDH1_cn	10,8080078
CDK4_cn	19,21333122	CDK6_cn	-3,404042328	CDKN2A_wt	122,007447
CDKN2A_cn	134,2255769	CDKN2C_wt	1,373084885	CDKN2C_cn	12,36115768
CDKN2a.p14._wt	153,7457211	CTNNB1_wt	25,17669328	CTNNB1_cn	14,0477282
CYLD_wt	4,067468675	EGFR_wt	7,44652644	EGFR_cn	0,889132715
EP300_cn	14,18503337	ERBB2_wt	7,108732684	ERBB2_cn	55,78107205
EZH2_wt	-12,3794173	EZH2_cn	-1,970967279	FAM123B_wt	10,5668683
FAM123B_cn	12,25838892	FBXW7_wt	30,18933049	FBXW7_cn	6,295771371
FGFR2_cn	30,10501665	FGFR3_wt	9,098347722	FGFR3_cn	9,414153789
FLCN_wt	5,47104567	FLT3_wt	4,051218248	FLT3_cn	-0,848579219
GNAS_wt	16,27211687	GNAS_cn	17,4989612	HRAS_wt	-39,21435245
IDH1_wt	12,32566845	IDH1_cn	8,616086164	JAK2_wt	9,264806901
JAK2_cn	-0,242646969	KDM5C_wt	6,926119405	KDM5C_cn	2,686800292
KDM6A_wt	16,58677504	KDM6A_cn	4,005536184	KDR_cn	7,751043029
KIT_cn	13,33010909	KRAS_wt	111,5493929	KRAS_cn	-1,450278417
MAP2K4_wt	14,49869672	MAP2K4_cn	3,158561346	MDM2_cn	29,3888009
MET_cn	18,37416265	MLH1_wt	61,49071831	MLH1_cn	3,453643234
MLLT3_cn	-26,64515808	MSH2_wt	-2,030550553	MSH2_cn	4,049141974
MSH6_wt	-0,430785072	MSH6_cn	4,448594341	MYC_cn	40,05777068
MYCL1_cn	3,246714122	MYCN_cn	33,67147491	NF1_wt	29,65187136
NF1_cn	4,887243319	NF2_wt	40,1714892	NF2_cn	19,47629403
NOTCH1_wt	48,05918449	NRAS_wt	60,86243961	NRAS_cn	19,17492398
PDGFRA_cn	-15,84609699	PIK3CA_wt	46,09398915	PIK3CA_cn	22,99938931
PIK3R1_wt	30,78934492	PIK3R1_cn	15,10361112	PTEN_wt	66,63034621
PTEN_cn	55,42201737	RB1_wt	33,50767448	RB1_cn	5,313157474
RUNX1_wt	9,562460173	SETD2_wt	-13,7016384	SMAD4_wt	125,2284457
SMAD4_cn	1,90511325	SMARCA4_wt	11,93739617	SMARCA4_cn	-3,844929231
SMARCB1_cn	1,322720248	SMO_cn	-5,067195894	SOCS1_cn	0,979144347
STK11_wt	41,09522305	STK11_cn	-2,391314291	TET2_wt	39,48124438
TP53_wt	66,76338137	TP53_cn	-1,242357163	TSC1_wt	25,11899834
TSC1_cn	19,69118403	VHL_wt	18,33238528	VHL_cn	14,12852076
WT1_cn	7,969079223	PubchemFP0	5,035468619	PubchemFP1	0,752521081
PubchemFP2	-0,727583648	PubchemFP6	9,778226521	PubchemFP9	1,50941988
PubchemFP10	1,572153323	PubchemFP11	1,64273151	PubchemFP12	2,038613269
PubchemFP13	0,52464496	PubchemFP14	2,089585921	PubchemFP15	4,119376236
PubchemFP16	3,416851802	PubchemFP17	2,251364586	PubchemFP18	2,651439991
PubchemFP19	6,334751122	PubchemFP20	7,526925112	PubchemFP21	1,405399459
PubchemFP22	4,249062984	PubchemFP23	6,340606195	PubchemFP24	4,347539016
PubchemFP25	2,627703716	PubchemFP30	1,5750988	PubchemFP33	3,588704046
PubchemFP34	4,340509007	PubchemFP37	4,735557781	PubchemFP38	4,393460716
PubchemFP43	2,092036219	PubchemFP44	1,670459932	PubchemFP46	3,490975079
PubchemFP93	4,167294084	PubchemFP115	1,898937231	PubchemFP116	2,125094311
PubchemFP117	0,23584835	PubchemFP118	0,770053538	PubchemFP129	0,673594341
PubchemFP130	0,681782642	PubchemFP132	0,112819959	PubchemFP143	4,447260711
PubchemFP144	1,977295207	PubchemFP145	4,988148136	PubchemFP146	4,298797526
PubchemFP147	-2,492472564	PubchemFP148	3,421945192	PubchemFP149	3,346100388
PubchemFP150	3,953383976	PubchemFP152	3,961091716	PubchemFP153	2,943305466
PubchemFP155	5,762142508	PubchemFP156	5,790322808	PubchemFP157	6,137785825
PubchemFP159	1,658986287	PubchemFP160	1,765575423	PubchemFP164	1,090568457
PubchemFP167	0,692580181	PubchemFP178	13,09839813	PubchemFP179	1,830746696
PubchemFP180	5,543624655	PubchemFP181	3,841755281	PubchemFP182	3,607603577
PubchemFP183	2,182317658	PubchemFP184	2,62037662	PubchemFP185	1,505261329
PubchemFP186	2,912598029	PubchemFP187	15,92119099	PubchemFP188	10,58199805
PubchemFP189	0,854985581	PubchemFP190	-3,304221549	PubchemFP191	2,359198788
PubchemFP192	4,007219184	PubchemFP193	5,363962259	PubchemFP194	5,856501046
PubchemFP195	2,928128932	PubchemFP199	2,932757531	PubchemFP200	2,112327747
PubchemFP206	5,798364759	PubchemFP213	-2,070713526	PubchemFP214	-3,392875085
PubchemFP218	1,375421627	PubchemFP219	1,437694656	PubchemFP227	-2,205312332
PubchemFP228	0,327070955	PubchemFP232	-2,337298567	PubchemFP233	-3,039825938
PubchemFP241	0,580965068	PubchemFP246	0,587985031	PubchemFP247	0,557462073
PubchemFP248	1,138195392	PubchemFP252	0,281979609	PubchemFP255	2,183031488
PubchemFP256	3,200082677	PubchemFP257	3,239240945	PubchemFP258	3,581859076

PubchemFP259	3,645707494	PubchemFP260	5,383852081	PubchemFP261	3,396994653
PubchemFP262	0,260697604	PubchemFP274	9,489842924	PubchemFP276	9,73303866
PubchemFP283	1,699943936	PubchemFP284	2,226823221	PubchemFP285	2,754361495
PubchemFP286	3,142872114	PubchemFP287	6,042216824	PubchemFP293	3,40661217
PubchemFP294	4,424720337	PubchemFP297	2,163793107	PubchemFP298	3,52476922
PubchemFP299	4,247559907	PubchemFP300	3,57026757	PubchemFP301	9,882203823
PubchemFP305	1,755239884	PubchemFP308	7,59258311	PubchemFP314	1,807685891
PubchemFP327	1,789212726	PubchemFP328	2,029866309	PubchemFP330	1,99003397
PubchemFP332	1,347041709	PubchemFP333	2,245755768	PubchemFP334	5,280809379
PubchemFP335	3,262431726	PubchemFP336	3,417376971	PubchemFP337	21,26925714
PubchemFP338	3,212313392	PubchemFP339	3,706597791	PubchemFP340	3,050818127
PubchemFP341	5,050882829	PubchemFP342	4,500048821	PubchemFP344	2,114378594
PubchemFP345	4,429391664	PubchemFP346	7,275694547	PubchemFP347	3,064296016
PubchemFP349	6,596801925	PubchemFP350	3,79495162	PubchemFP351	3,107698639
PubchemFP352	3,161583937	PubchemFP353	3,428715698	PubchemFP355	2,22820657
PubchemFP356	2,707268388	PubchemFP357	3,528529106	PubchemFP358	3,240715875
PubchemFP359	3,189534043	PubchemFP360	-0,073460272	PubchemFP362	4,759589312
PubchemFP363	6,345789851	PubchemFP364	5,404394126	PubchemFP365	3,301888238
PubchemFP366	4,403417199	PubchemFP367	2,811453697	PubchemFP368	3,687785375
PubchemFP370	1,62694668	PubchemFP371	0,896152142	PubchemFP372	4,100322872
PubchemFP373	2,544290732	PubchemFP374	6,524251158	PubchemFP375	5,388125927
PubchemFP376	2,253476728	PubchemFP377	3,998904935	PubchemFP378	3,012664359
PubchemFP379	1,739538989	PubchemFP380	2,088844879	PubchemFP381	2,395865041
PubchemFP382	2,92243145	PubchemFP383	1,784338744	PubchemFP384	1,981711201
PubchemFP385	3,321384179	PubchemFP386	4,608848331	PubchemFP387	3,124193313
PubchemFP388	2,24370451	PubchemFP389	2,150843379	PubchemFP390	2,177511627
PubchemFP391	4,359004486	PubchemFP392	2,871800257	PubchemFP393	4,35114589
PubchemFP394	3,040897483	PubchemFP395	9,441341356	PubchemFP396	3,241731305
PubchemFP397	2,658294356	PubchemFP398	2,72160104	PubchemFP399	1,945200372
PubchemFP400	1,750086581	PubchemFP403	2,264048303	PubchemFP404	0,018549222
PubchemFP405	3,514824917	PubchemFP406	3,795958968	PubchemFP407	1,640161
PubchemFP408	5,154791068	PubchemFP409	0,412641653	PubchemFP411	1,507783558
PubchemFP412	2,748535201	PubchemFP413	2,20760982	PubchemFP414	2,884584666
PubchemFP416	1,674767359	PubchemFP417	1,811059695	PubchemFP418	3,677145081
PubchemFP419	3,126702642	PubchemFP420	3,100899807	PubchemFP421	3,423907832
PubchemFP422	3,203204277	PubchemFP423	1,051853582	PubchemFP425	1,46185431
PubchemFP427	1,759641327	PubchemFP428	1,846426958	PubchemFP429	3,544931135
PubchemFP430	3,379433801	PubchemFP431	3,985399855	PubchemFP432	4,547138317
PubchemFP434	4,955938618	PubchemFP435	3,824468546	PubchemFP436	1,406668381
PubchemFP437	3,087977178	PubchemFP438	3,328030389	PubchemFP439	5,098811244
PubchemFP440	3,371711741	PubchemFP441	1,641441249	PubchemFP442	2,830078421
PubchemFP443	4,124533085	PubchemFP445	4,235340814	PubchemFP446	7,803190299
PubchemFP447	3,006462006	PubchemFP448	1,51904661	PubchemFP449	3,147487724
PubchemFP450	4,426407905	PubchemFP451	4,313408275	PubchemFP452	4,296891352
PubchemFP453	3,892483043	PubchemFP454	1,32564735	PubchemFP456	1,603232268
PubchemFP457	2,417281741	PubchemFP458	4,596268258	PubchemFP459	2,105333938
PubchemFP460	1,884428977	PubchemFP461	2,60105	PubchemFP462	3,452725035
PubchemFP464	2,731212656	PubchemFP465	2,697543271	PubchemFP466	3,545236737
PubchemFP467	2,582009855	PubchemFP469	-3,398708243	PubchemFP470	1,017866822
PubchemFP471	3,165212358	PubchemFP472	2,661769675	PubchemFP473	-0,48248963
PubchemFP474	1,517183109	PubchemFP475	-0,354555995	PubchemFP476	2,867624807
PubchemFP477	3,278652607	PubchemFP480	2,671244312	PubchemFP481	1,498522939
PubchemFP482	2,935790523	PubchemFP483	0,515319295	PubchemFP484	1,920114054
PubchemFP485	4,25205125	PubchemFP486	1,370411975	PubchemFP487	4,065647853
PubchemFP489	3,084472626	PubchemFP490	0,580796968	PubchemFP493	4,595630752
PubchemFP494	3,143002119	PubchemFP495	2,073245766	PubchemFP497	2,378332543
PubchemFP498	4,205818662	PubchemFP499	3,476337074	PubchemFP500	3,021393908
PubchemFP501	3,386302732	PubchemFP504	2,14597523	PubchemFP505	0,874517483
PubchemFP506	2,585141176	PubchemFP507	3,147213064	PubchemFP508	3,89216495
PubchemFP509	1,749937259	PubchemFP514	10,8578101	PubchemFP515	4,257585381
PubchemFP516	2,762124916	PubchemFP517	1,351071252	PubchemFP518	0,918565527
PubchemFP519	3,279457996	PubchemFP520	0,086609553	PubchemFP521	1,344152939
PubchemFP523	3,09379198	PubchemFP524	0,68437173	PubchemFP530	2,950498183
PubchemFP531	2,311640845	PubchemFP533	3,484301284	PubchemFP534	2,957503756
PubchemFP535	5,062451829	PubchemFP536	5,534665206	PubchemFP537	3,459472982
PubchemFP538	3,479138091	PubchemFP539	-0,43274116	PubchemFP540	7,264695098
PubchemFP541	2,704938307	PubchemFP543	5,030502599	PubchemFP544	1,966940486
PubchemFP545	1,918594802	PubchemFP547	3,650932127	PubchemFP548	8,931670877
PubchemFP549	1,876087966	PubchemFP550	3,307272233	PubchemFP551	0,809790144
PubchemFP552	1,078056982	PubchemFP553	4,700722058	PubchemFP554	2,119269061
PubchemFP555	2,925741993	PubchemFP556	0,826860807	PubchemFP558	0,74917847
PubchemFP560	5,110426481	PubchemFP563	11,69008325	PubchemFP564	0,080968193
PubchemFP565	2,517460403	PubchemFP566	5,637144883	PubchemFP567	4,402790198
PubchemFP568	3,681791436	PubchemFP569	2,561367378	PubchemFP570	1,887455511
PubchemFP572	2,327700357	PubchemFP573	3,103585815	PubchemFP574	8,637006738
PubchemFP575	3,906413347	PubchemFP577	3,434280609	PubchemFP578	1,253430472
PubchemFP579	4,974276126	PubchemFP580	4,350825551	PubchemFP581	27,9611602
PubchemFP582	1,681736888	PubchemFP583	1,302077708	PubchemFP584	0,604363524
PubchemFP585	1,917250968	PubchemFP586	3,681558892	PubchemFP588	1,779014021
PubchemFP589	2,311423097	PubchemFP591	3,545725403	PubchemFP592	2,227913207
PubchemFP593	3,692339994	PubchemFP594	2,738371646	PubchemFP595	2,030224358
PubchemFP596	2,640274331	PubchemFP597	3,748082056	PubchemFP598	3,2822545
PubchemFP599	2,770113185	PubchemFP600	2,6002304	PubchemFP601	2,926047837
PubchemFP602	5,340688288	PubchemFP603	1,083790777	PubchemFP604	8,873180736
PubchemFP605	2,998808254	PubchemFP606	2,776423494	PubchemFP607	1,535273389
PubchemFP608	2,045482412	PubchemFP609	3,68854547	PubchemFP610	2,766570366
PubchemFP611	3,220448873	PubchemFP612	3,308343548	PubchemFP613	2,864733154
PubchemFP614	3,900336632	PubchemFP615	4,052481121	PubchemFP616	3,214439179
PubchemFP618	0,892978037	PubchemFP619	2,386827311	PubchemFP620	2,093764106
PubchemFP621	2,083110694	PubchemFP622	-2,211401409	PubchemFP623	3,864328125
PubchemFP624	4,178651183	PubchemFP625	3,847376552	PubchemFP626	2,638002847
PubchemFP628	1,824278915	PubchemFP629	2,424761452	PubchemFP630	1,706165914
PubchemFP632	2,284462642	PubchemFP633	3,793120737	PubchemFP634	1,112102323

PubchemFP636	3,42001581	PubchemFP637	3,579524692	PubchemFP638	3,720332063
PubchemFP639	2,084604434	PubchemFP640	0,834208394	PubchemFP641	3,307970637
PubchemFP642	28,37317767	PubchemFP644	2,465159236	PubchemFP645	6,003074486
PubchemFP646	2,043651855	PubchemFP647	2,320996738	PubchemFP648	0,089614742
PubchemFP650	2,137634745	PubchemFP651	2,623343775	PubchemFP652	2,348846841
PubchemFP653	2,670743772	PubchemFP654	3,54665693	PubchemFP655	3,96546641
PubchemFP656	3,550307079	PubchemFP657	1,967683655	PubchemFP658	5,60002326
PubchemFP659	5,616627598	PubchemFP660	-0,71142259	PubchemFP661	12,16216874
PubchemFP662	2,378482271	PubchemFP664	-0,272578204	PubchemFP665	2,974827701
PubchemFP666	2,818688973	PubchemFP668	1,044766523	PubchemFP669	1,549215762
PubchemFP670	1,831835136	PubchemFP671	4,7873435	PubchemFP673	4,922298182
PubchemFP674	3,275024502	PubchemFP675	2,176488304	PubchemFP676	3,322188622
PubchemFP677	1,008182326	PubchemFP678	0,184241405	PubchemFP679	1,549807243
PubchemFP680	2,533810793	PubchemFP681	6,402521434	PubchemFP682	6,002507529
PubchemFP683	3,055269419	PubchemFP684	5,09083814	PubchemFP685	2,997319401
PubchemFP686	8,749907884	PubchemFP687	7,384540765	PubchemFP688	5,828494223
PubchemFP689	3,299984335	PubchemFP690	4,229656106	PubchemFP691	3,336823487
PubchemFP692	5,498645003	PubchemFP693	6,031396707	PubchemFP694	1,395811294
PubchemFP695	4,289434475	PubchemFP696	4,024892491	PubchemFP697	7,863101091
PubchemFP698	6,635525106	PubchemFP699	4,137553971	PubchemFP700	3,391746725
PubchemFP701	3,707770987	PubchemFP702	3,575388058	PubchemFP703	2,984307521
PubchemFP704	5,069526116	PubchemFP705	3,508552305	PubchemFP706	-1,823278502
PubchemFP707	4,485150787	PubchemFP708	2,362951465	PubchemFP709	3,251570533
PubchemFP710	3,575177439	PubchemFP711	2,794713358	PubchemFP712	3,233427353
PubchemFP713	5,40347963	PubchemFP714	2,02592958	PubchemFP715	1,950547388
PubchemFP716	5,237212708	PubchemFP717	1,942742226	PubchemFP719	2,823975905
PubchemFP721	2,931304177	PubchemFP722	1,692648579	PubchemFP725	0,118438053
PubchemFP728	5,819371228	PubchemFP729	1,872413923	PubchemFP733	1,629328477
PubchemFP734	5,337082078	PubchemFP735	3,521009467	PubchemFP736	1,9599078
PubchemFP737	2,975633098	PubchemFP738	3,522152687	PubchemFP740	2,909986028
PubchemFP742	1,962774021	PubchemFP743	1,390092533	PubchemFP745	0,615838987
PubchemFP746	1,94741773	PubchemFP747	1,443646611	PubchemFP749	2,783839648
PubchemFP750	6,965900675	PubchemFP751	1,644319352	PubchemFP752	1,971222611
PubchemFP755	4,509168769	PubchemFP756	2,309154492	PubchemFP757	3,356086623
PubchemFP758	3,272900182	PubchemFP759	2,126912761	PubchemFP761	1,727962529
PubchemFP762	-2,788624266	PubchemFP763	4,306096212	PubchemFP764	0,797468078
PubchemFP766	1,610017478	PubchemFP767	1,993364349	PubchemFP770	2,175669319
PubchemFP771	5,159418338	PubchemFP772	1,729336143	PubchemFP776	3,574076373
PubchemFP777	2,316310951	PubchemFP778	1,98655155	PubchemFP779	5,146010139
PubchemFP780	2,161756508	PubchemFP782	2,165292143	PubchemFP784	3,109873155
PubchemFP785	2,080134819	PubchemFP788	-0,065853004	PubchemFP791	6,09331101
PubchemFP792	1,884602673	PubchemFP796	2,129760995	PubchemFP797	3,383468658
PubchemFP798	3,729518579	PubchemFP799	1,880799198	PubchemFP800	3,023748874
PubchemFP801	3,818265185	PubchemFP803	3,299252876	PubchemFP805	1,964405158
PubchemFP806	1,157924779	PubchemFP808	0,486156896	PubchemFP809	1,93892923
PubchemFP810	1,472323875	PubchemFP812	2,807882462	PubchemFP813	6,900219219
PubchemFP814	1,900581668	PubchemFP815	1,874921751	PubchemFP818	5,12807583
PubchemFP819	3,801071966	PubchemFP820	2,853649623	PubchemFP821	3,292301438
PubchemFP822	2,303791526	PubchemFP824	2,331609804	PubchemFP825	-2,082496566
PubchemFP826	4,276495838	PubchemFP827	1,461280819	PubchemFP829	1,838977891
PubchemFP830	2,210479593	PubchemFP833	2,548097339	PubchemFP834	5,203282609
PubchemFP835	1,945918279	PubchemFP839	1,337134286	PubchemFP840	-2,583902923
PubchemFP860	-2,347280754	PubchemFP861	-2,46705644	nAcid	3,332081242
apol	7,397373918	naAromAtom	11,74694985	nAromBond	11,4684336
nAtom	7,2529769	nHeavyAtom	5,259682055	nH	16,39710999
nB	9,575708041	nC	11,12802273	nN	8,839979384
nO	9,638108827	nS	5,585545547	nP	1,710920882
nF	8,47355966	nCl	6,70769428	nBr	1,780929151
nI	3,619785573	ATSc1	11,72103653	ATSc2	17,8538851
ATSc3	28,54495272	ATSc4	19,96408713	ATSc5	20,06291788
ATSm1	11,21852329	ATSm2	7,44021864	ATSm3	5,82230879
ATSm4	9,923189169	ATSm5	9,901599117	ATSp1	7,459706434
ATSp2	7,60387196	ATSp3	8,5166431	ATSp4	10,31620231
ATSp5	9,085094685	nBase	9,259139652	nBonds	5,539070583
nBonds2	8,839371577	nBondsS	7,842189569	nBondsS2	8,108908035
nBondsS3	9,394517358	nBondsD	10,11982319	nBondsD2	10,51460108
nBondsT	3,415922543	bpol	7,499026382	C1SP1	3,314858487
C2SP1	2,060077231	C1SP2	10,80477492	C2SP2	15,89893105
C3SP2	15,58708099	C1SP3	11,80770938	C2SP3	11,30300554
C3SP3	6,873386489	C4SP3	6,04992395	SCH.3	2,507347999
SCH.4	6,469188021	SCH.5	13,14783506	SCH.6	17,96205057
SCH.7	15,82997014	VCH.3	2,719564999	VCH.4	7,303959982
VCH.5	15,19849046	VCH.6	12,87316705	VCH.7	23,55994181
SC.3	12,04145508	SC.4	14,00599383	SC.5	7,383570865
SC.6	11,8610675	VC.3	10,83903348	VC.4	21,84588415
VC.5	14,93723049	VC.6	11,37794061	SPC.4	9,580457638
SPC.5	13,07353486	SPC.6	9,08862701	VPC.4	9,638926115
VPC.5	9,670087183	VPC.6	9,535915933	ECCEN	9,649229452
fragC	14,87946664	nHBAcc	7,168215635	nHBAcc2	7,134828175
nHBAcc3	16,35565225	nHBAcc_Lipinski	6,503030978	nHBDon	31,42144542
nHBDon_Lipinski	9,664117003	nAtomLC	11,80361895	nAtomP	11,39901697
nAtomLAC	12,64567921	MLogP	23,40030942	McGowan_Volume	9,279004335
MDEC.11	7,888492134	MDEC.12	11,00859844	MDEC.13	13,65455131
MDEC.14	10,42887048	MDEC.22	18,30938998	MDEC.23	10,69907753
MDEC.24	13,69904338	MDEC.33	17,65236065	MDEC.34	16,21610379
MDEC.44	4,247316136	MDEO.11	18,54133766	MDEO.12	38,35723776
MDEO.22	5,161496335	MDEN.11	9,259706793	MDEN.12	8,763776733
MDEN.13	28,42443718	MDEN.22	13,40163382	MDEN.23	8,311805906
MDEN.33	9,272423975	MLFER_A	12,54818716	MLFER_BH	13,26045329
MLFER_BO	18,27897289	MLFER_S	13,46270908	MLFER_E	14,21338674
MLFER_L	22,05257189	PetitjeanNumber	13,81155272	nRing	4,687561127
n3Ring	1,821278134	n4Ring	0,948640063	n5Ring	7,993345974
n6Ring	13,20644912	n7Ring	-0,434223307	n8Ring	-2,944398907

n9Ring	-0,181328412	n10Ring	1,26287563	nG12Ring	1,980471213
nFRing	7,77362455	nF6Ring	0,339097763	nF8Ring	0,761810914
nF9Ring	6,584906595	nF10Ring	8,550225307	nF11Ring	2,192974805
nF12Ring	3,0718827	nFG12Ring	4,769270186	nTRing	7,905911957
nT4Ring	0,937080721	nT5Ring	7,967636357	nT6Ring	13,1341923
nT7Ring	-0,757553318	nT8Ring	1,720740882	nT9Ring	6,755727742
nT10Ring	8,717981457	nT11Ring	2,168133585	nT12Ring	3,227169333
nTG12Ring	4,868991532	nRotB	11,34629835	TopoPSA	15,2831086
VAdjMat	5,539727381	MW	12,34151936	WTPT.1	7,35349285
WTPT.2	12,98955477	WTPT.3	10,86242205	WTPT.4	16,41680473
WTPT.5	14,67650623	WPATH	16,3545136	WPOL	8,653764565
Zagreb	8,046209245	X533	2,689047064	X131	6,043162967
X486	2,194357265	X93	3,629466182	X28	3,714148992
X282	2,932288034	X429	2,816670947	X292	2,784258921
X716	2,679971733	X893	2,180345559	X564	3,658207578
X490	2,34918716	X132	7,083041069	X442	4,207258425
X327	4,542205223	X90	2,521110405	X333	6,868836084
X562	3,219728746	X51	3,085486643	X390	7,126369427
X693	4,629926703	X346	6,827906131	X686	3,072518304
X779	3,02187265	X27	2,909601895	X569	5,04697968
X687	2,372636581	X128	5,605741997	X218	5,656526873
X430	2,508314463	X859	2,677560743	X415	5,129922605
X541	3,649699997	X269	3,640523129	X898	5,667581821
X258	3,076320528	X13	3,854120359	X384	4,456409957
X409	4,088032907	X580	2,777753994	X263	4,221398588
X426	4,064737459	X531	2,780034551	X575	3,133133375
X519	2,577781002	X189	2,158220581	X553	3,108110035
X818	4,340217861	X559	7,081057781	X407	4,947733202
X37	3,017472932	X708	4,736635964	X462	3,729512377
X764	2,647017612	X681	2,543338429	X160	3,235457006
X222	3,791156354	X45	6,969980206	X735	3,740496858
X800	2,707683438	X100	2,129745411	X416	3,240349902
X444	4,610706093	X107	5,064144649	X115	3,225509011
X913	3,023999989	X599	4,203985076	X241	4,174981356
X202	5,015508259	X551	3,504476424	X560	2,87391391
X361	5,226020953	X583	3,442226621	X493	2,644005181
X1	1,944626606	X5	4,974526286	X58	4,056335883
X249	4,594272575	X641	3,77309247	X141	3,558112952
X809	3,980875158	X793	0,465929656	X633	2,932972595
X622	2,665660401	X59	16,47605452	X577	2,623446072
X875	4,36391711	X145	3,736249754	X613	4,663459656
X827	2,709032745	X607	3,689911713	X871	3,806570926
X592	2,55192084	X423	4,799355507	X322	3,126660535
X495	2,983664063	X658	2,580888199	X369	3,593170963
X804	3,340600555	X103	2,441607061	X268	3,629364244
X826	4,995576544	X122	2,830337231	X422	3,769586012
X452	2,515616971	X363	2,946411007	X579	2,291786064
X909	3,91543789	X786	3,574552504	X568	4,522787221
X769	2,900454829	X840	2,945448536	X889	3,261406163
X825	3,382508811	X75	2,745853492	X555	9,956459772
X498	3,535725447	X38	3,44731556	X167	5,519371036
X480	2,736281389	X275	5,227795285	X783	3,176227661
X19	2,340273676	X801	3,132050459	X114	4,240445921
X270	2,980343527	X54	4,715852152	X405	4,072777026
X849	3,011271487	X304	3,855850168	X299	2,995618905
X731	2,683201233	X193	10,92432422	X532	2,399844935
X198	3,158618236	X710	2,36097675	X417	4,954457077
X305	4,376629625	X272	5,021129076	X721	4,4005475
X864	4,635029204	X703	3,788200327	X732	3,323586499
X391	4,65549171	X759	2,494151998	X97	2,523903747
X497	2,650877703	X459	2,827014055	X311	4,532187412
X905	2,688523664	X335	2,385905701	X180	3,950908682
X18	6,738851395	X887	4,069302286	X744	3,080414486
X211	2,78796696	X205	4,014581693	X91	2,594762671
X356	4,015826449	X194	4,074340973	X830	2,996632668
X799	3,994560906	X520	2,469545074	X352	4,616571089
X861	3,112967316	X184	3,346007416	X509	2,311182947
X223	2,771326145	X22	3,511676346	X8	2,726095979
X762	2,791248083	X637	2,466434058	X829	4,881767784
X159	2,780708667	X576	2,51532598	X790	4,495072022
X892	3,13642764	X665	2,525685216	X885	5,457031425
X48	2,468495141	X788	2,439646512	X803	3,941464909
X379	2,684597367	X432	3,433967214	X798	4,02791627
X770	3,176729383	X228	3,50510575	X257	3,219861336
X844	3,622368283	X71	3,139157078	X860	2,937426203
X171	3,444339492	X726	3,183132321	X151	4,212457287
X857	3,24071417	X886	4,399191421	X868	3,279251388
X872	2,323844849	X567	3,059196583	X220	3,145317781
X40	8,470149559	X341	4,848236814	X910	4,727220666
X736	4,821455029	X440	6,75685441	X877	2,997863238
X50	3,32658302	X660	3,110187804	X154	3,210621177
X719	2,703335535	X673	3,222316132	X169	3,173494362
X745	6,196082921	X146	3,920771395	X791	1,854921906
X425	3,266740542	X454	2,961292974	X517	2,668111988
X604	2,720535615	X645	3,365581639	X108	3,13534011
X904	3,349401575	X620	2,566311859	X420	3,861963174
X661	3,087384224	X226	2,957510804	X571	3,118509017
X362	2,477615187	X153	2,46873442	X806	2,502745902
X915	2,58935681	X276	5,085615334	X204	3,443713818
X178	2,912510098	X183	3,286648737	X364	3,224758993
X315	9,240138069	X496	4,390752217	X526	3,3113364
X436	3,589499828	X513	2,260605764	X300	3,882119994
X460	2,634503898	X31	5,774634076	X891	2,53472356

X469	2,749811947	X853	4,179577519	X324	3,797532131
X743	3,236892375	X651	2,727355791	X89	2,480529072
X181	3,879926472	X49	4,472035446	X15	4,539076189
X366	7,136719308	X457	3,110101945	X695	2,969079594
X371	4,666421933	X113	2,813135562	X503	2,713593724
X663	5,755929627	X834	3,737314991	X212	2,662243495
X856	2,973607357	X284	5,981096986	X808	2,546903689
X56	5,335878654	X224	4,116188345	X357	2,552824104
X862	3,185480196	X488	2,155385633	X96	3,540507257
X612	5,006468061	X802	3,461150849	X157	2,555476516
X707	3,091511772	X724	4,319012615	X574	2,221793984
X129	3,46854724	X466	2,510998479	X771	3,899617344
X547	2,890280634	X676	3,885309188	X372	3,194753576
X267	4,087640484	X6	3,503700394	X510	2,031179526
X652	2,975560033	X265	10,98824577	X487	2,56333188
X155	2,498893047	X494	2,703782759	X590	4,113435018
X881	3,556294205	X754	2,417877961	X147	2,763305087
X427	2,281110876	X655	3,504647365	X370	4,108362136
X884	1,996136171	X846	3,105157532	X833	3,660816228
X593	3,04202858	X698	3,259984818	X441	3,964314536
X377	4,606312955	X235	3,566278078	X404	3,87006467
X539	2,374446376	X717	3,886049845	X631	2,764723063
X614	5,83255618	X101	3,438431626	X691	5,812366025
X439	4,235180486	X843	3,348522353	X749	2,380025092
X25	3,088495325	X761	2,750548396	X899	3,811522201
X774	3,281936682	X820	3,69149935	X289	2,215833178
X136	3,072503814	X476	2,289342756	X336	2,971453555
X483	2,905115761	X26	3,632504973	X83	2,319676563
X456	4,117866778	X505	2,902759917	X260	6,388815856
X144	3,31990361	X174	2,63075613	X368	2,447030247
X634	2,730569204	X453	2,347998968	X643	3,150496472
X545	3,955163354	X246	3,194773259	X540	2,239068124
X400	7,036250361	X296	3,033558984	X321	1,952629168
X112	4,108604605	X85	2,532657743	X720	7,777442278
X435	4,301274028	X250	3,960018258	X582	2,565115843
X360	3,117790793	X831	4,627194068	X254	3,815207661
X474	2,517831672	X669	4,127674677	X374	4,079633018
X342	3,800106978	X323	3,435175266	X624	3,393421075
X124	2,899127644	X527	2,777063339	X572	5,72964678
X475	2,239081419	X556	3,482354828	X216	3,981100478
X240	3,721422181	X399	4,370984852	X331	6,520776183
X566	4,579031856	X598	4,392444041	X699	2,551250526
X521	3,685002201	X588	3,354791305	X210	2,878691078
X753	2,412914948	X878	3,545874497	X756	2,713620154
X741	2,824426696	X882	3,06341744	X102	4,248268374
X611	2,586483484	X679	4,276386046	X256	3,365696324
X478	2,526960398	X344	3,957795559	X740	1,671548371
X229	5,149180935	X383	5,074108433	X773	3,419300761
X35	2,374885691	X775	2,730948353	X838	3,838182774
X565	3,493091888	X0	1,497395152	X455	3,899635522
X586	2,144647541	X69	4,523672613	X355	2,757366835
X388	3,856590647	X814	3,149862101	X748	2,950916848
X227	3,62948851	X778	3,236247678	X21	3,550076037
X401	11,6840971	X796	3,536112065	X570	3,078290374
X373	4,068183922	X408	3,993568998	X819	3,002318339
X601	3,808689288	X62	4,944435765	X403	5,005342165
X421	3,351519268	X578	3,121484521	X4	5,407242557
X365	4,362742815	X528	2,749722486	X133	4,739913803
X714	3,720441266	X608	5,79388447	X851	1,947519588
X522	2,27151853	X412	3,77953925	X14	3,521730058
X280	3,773857613	X697	2,948367922	X84	4,020389431
X176	2,772708834	X332	6,593052101	X727	3,696560188
X777	3,354606777	X696	2,750037391	X789	4,370963139
X646	2,378820436	X689	4,24546232	X319	3,202941973
X126	3,610487914	X111	9,684872842	X890	2,948661841
X758	2,591851386	X534	3,255715816	X53	3,177900383
X308	6,088776591	X448	3,42168043	X367	2,708195847
X301	5,80778538	X589	3,571720711	X434	3,604145057
X894	5,300973999	X671	3,064578169	X516	5,831171602
X591	3,243469461	X837	4,210887264	X461	4,766532999
X780	3,219844278	X92	5,514237162	X117	2,445364139
X99	2,124450637	X60	3,75792382	X847	3,478023253
X314	5,778057015	X606	2,861860938	X842	2,669483844
X350	3,253400696	X243	5,562273558	X538	2,368145865
X471	2,209659568	X203	3,72817277	X595	2,591383262
X867	3,485031216	X805	4,07907645	X162	3,428091022
X874	5,544028603	X836	3,890539366	X152	3,901786718
X479	2,298060165	X668	4,233096106	X512	2,758251617
X546	2,831551083	X419	4,118563977	X221	2,6111164
X33	11,27343702	X738	2,561186912	X502	2,436646444
X207	2,926854943	X130	5,003508725	X120	3,311803519
X561	2,841406221	X596	3,061964074	X278	6,677045331
X86	2,212189099	X458	3,976452605	X135	4,579024698
X382	3,020968024	X206	3,499518928	X307	8,730730774
X137	4,119310921	X47	4,197246297	X468	3,410002387
X737	3,775760328	X543	3,568661305	X537	2,895530101
X438	4,165108518	X16	4,023766692	X316	2,853112285
X42	5,112505167	X392	2,874073579	X597	4,475917146
X573	3,784940826	X616	2,772284374	X811	3,609278462
X558	5,658484952	X855	2,877467283	X584	2,046289507
X290	2,330103236	X73	3,373180554	X850	2,242081436
X303	2,97633267	X381	3,474993806	X196	2,903540869
X199	3,252871162	X845	2,719259903	X626	3,66766951

X402	2,545485933	X659	7,175326048	X179	3,887994648
X654	3,891917633	X870	3,085676013	X896	2,28289105
X623	3,659819191	X914	2,910087085	X127	3,893975714
X835	2,045784136	X328	2,988707725	X628	4,080891343
X340	3,078619959	X214	3,547696429	X233	2,79980451
X286	3,525299022	X345	2,653244604	X883	1,705356694
X482	3,312790186	X832	3,245122213	X649	2,976902638
X902	4,237953743	X734	3,527720034	X248	3,938069631
X76	4,495876205	X186	4,05605175	X279	3,455408658
X472	2,759004275	X666	2,468067583	X866	2,470004239
X329	3,555982804	X767	2,680742246	X785	3,336295962
X433	2,890544412	X118	2,017476378	X817	3,663341765
X44	4,82821855	X694	2,665674964	X752	4,359348674
X349	3,502838301	X67	6,6251809	X217	3,298598391
X863	3,594910661	X283	3,388935988	X713	4,839686371
X337	3,93156485	X766	2,839392649	X138	3,772175994
X262	3,235244738	X288	3,779103299	X358	4,112157089
X32	2,547079651	X879	2,576673987	X123	2,70456457
X389	4,764929235	X291	2,200541353	X215	3,520349017
X317	4,739901677	X554	3,412864107	X109	3,129044146
X106	6,768093704	X156	2,404871057	X841	3,325393942
X29	3,15995143	X852	4,014219717	X318	5,987959501
X353	3,516282855	X544	3,61502478	X348	3,646036226
X242	3,691029514	X166	5,591903101	X338	3,527916197
X201	4,931232765	X848	3,068683233	X281	5,503499383
X431	3,761082402	X293	2,384300334	X746	3,220365873
X680	2,17364781	X376	5,370459068	X393	3,360893393
X470	2,628295464	X712	2,305336005	X386	2,94636611
X715	6,755482105	X9	3,31261383	X310	3,350790125
X656	2,794389055	X177	2,945090724	X326	3,142928008
X81	7,656274651	X375	3,839923109	X742	3,32370589
X815	2,869974894	X763	2,594579821	X253	3,556739172
X20	2,282161295	X858	-0,099287679	X396	2,225400965
X418	4,195179418	X200	4,51145313	X149	3,49727543
X644	2,71461831	X822	2,519367524	X816	2,29032894
X163	3,436627457	X865	2,607011986	X23	3,045898683
X615	3,809319408	X185	4,160457034	X636	3,068494098
X916	2,381414491	X320	3,511104286	X705	3,675467781
X87	3,876489825	X230	3,667477771	X36	2,329502143
X24	2,973820888	X164	3,242043109	X895	6,741618599
X34	2,577620274	X273	3,271442342	X231	3,576490835
X445	10,71501927	X610	2,621243483	X380	3,700585544
X617	3,109337026	X43	4,408085563	X722	4,153864841
X7	2,850733759	X706	2,414336492	X797	3,852055378
X406	5,646045139	X140	3,285322286	X662	2,964338305
X94	4,618131143	X266	3,204438067	X869	3,534137846
X251	3,156702369	X190	2,508360656	X236	3,495408289
X339	2,412837554	X46	3,719177432	X912	1,773675447
X347	5,500468692	X747	3,869627065	X485	3,035153492
X675	2,690657706	X39	2,96710781	X312	4,944192771
X621	3,051824309	X261	3,794700623	X701	2,571543944
X464	2,153341501	X704	3,934657878	X451	4,475399857
X552	3,244560483	X529	2,4006832	X287	4,937626862
X600	4,075710886	X484	2,620628066	X667	2,498671311
X446	2,838786229	X603	3,850362636	X175	3,325368588
X134	3,486067827	X897	6,734342353	X295	2,222332546
X244	5,565704639	X670	5,243623415	X302	2,958248796
X625	4,047373994	X492	2,637411122	X121	2,549661269
X810	4,086959202	X648	2,245670126	X30	3,008216557
X511	2,479494722	X70	3,177041656	X491	2,544303427
X397	4,930029386	X57	6,723947797	X95	6,155027355
X168	3,911509092	X500	2,134849259	X765	3,086567478
X638	2,968775137	X506	2,804697069	X781	3,260990091
X602	3,933193402	X150	4,153514926	X682	2,663075416
X125	3,196999513	X594	5,222373202	X911	4,094140533
X225	2,889044972	X239	3,877252928	X192	1,880124712
X271	1,861960082	X394	2,893010874	X733	3,844878265
X182	3,961449946	X702	2,847936684	X709	2,330584681
X918	3,631028246	X585	2,000659	X750	3,956567233
X605	3,148960324	X550	3,534067619	X172	3,809972005
X143	2,935422183	X105	5,939052001	X787	4,724354006
X535	3,773846147	X255	3,841566492	X501	3,914479694
X755	2,57104101	X684	3,211268604	X813	3,441891772
X548	2,652120598	X2	1,742988097	X792	-0,654768194
X903	4,027234108	X351	3,442910845	X876	4,001626959
X395	5,375332449	X17	4,740838664	X473	2,472069119
X630	3,400415279	X443	3,880139417	X238	3,537616387
X306	5,187703233	X657	2,850552409	X247	3,401216147
X772	3,184440746	X839	3,758874733	X72	2,461611213
X812	3,752438017	X782	5,322749	X795	2,600630065
X187	3,461499081	X139	3,769187433	X690	4,974328657
X245	2,444992945	X807	3,121235199	X12	4,237847464
X74	4,073944673	X725	1,608854189	X188	1,893743541
X700	2,705385198	X692	2,926932332	X647	2,384226839
X82	3,236226721	X536	2,532117237	X294	2,366539831
X784	2,523571354	X518	3,30571065	X760	2,264566218
X465	2,386151598	X672	3,078648577	X907	3,837916958
X901	4,175359949	X285	5,957438395	X557	2,664903276
X219	6,310831132	X10	2,81584965	X674	2,814811602
X627	4,342414965	X88	4,459082976	X632	2,499284648
X908	6,396270638	X823	3,543899457	X313	4,856944892
X729	4,890202322	X232	3,627094014	X489	2,243748972
X78	3,134877137	X609	2,650022829	X387	4,377534083

X530	2,106519393	X264	4,396978685	X507	2,60290901
X80	3,820049004	X437	6,919411578	X678	4,705498876
X66	3,828169123	X508	2,728482169	X587	2,709595345
X161	4,41611935	X428	3,349896614	X650	2,96724215
X411	4,646934127	X104	2,366747946	X664	5,852769383
X524	2,752495333	X334	2,045201526	X685	8,081478168
X683	3,282334105	X325	4,667618549	X277	6,514638543
X110	5,342806208	X549	4,236975576	X424	3,809621427
X148	3,063205261	X514	2,401961129	X79	2,613096129
X619	3,126884402	X880	3,358138938	X259	2,716526934
X3	6,500663058	X776	3,428393283	X581	2,803904946
X41	6,462588211	X906	4,112510474	X635	2,518781416
X677	2,600264525	X919	5,329322448	X467	3,060254569
X450	2,877138402	X723	3,064849439	X449	3,41372239
X477	2,538887557	X64	2,716907456	X413	3,172948811
X173	2,716414864	X158	3,076842963	X525	2,434330858
X274	2,837083426	X234	4,170068147	X504	2,537162077
X116	3,02784115	X854	3,818041358	X98	2,323180853
X191	1,83134962	X688	3,691669095	X237	3,950278274
X252	3,828770315	X888	4,162262152	X119	4,209927633
X618	2,970663405	X213	3,755769507	X542	3,628299617
X165	3,892431836	X873	2,849401028	X523	4,01564681
X821	2,982564937	X718	2,566221544	X629	5,311617851
X354	3,580378535	X653	3,01043356	X208	3,382548405
X447	3,629212369	X170	3,849249032	X463	3,267575848
X642	3,23111047	X142	3,888315367	X730	3,575449908
X515	2,95294099	X343	5,308755236	X639	3,151953364
X65	4,314207254	X63	3,127518976	X11	3,869142988
X640	2,692514838	X209	2,705220793	X298	2,692445379
X330	4,549952853	X297	2,830117356	X824	1,784772204
X195	5,010873957	X900	15,77879541	X385	2,642611008
X414	5,450081611	X828	3,864505132	X751	3,26746318
X739	2,670088286	X55	2,760822073	X768	2,199370587
X68	2,632319294	X52	2,817392739	X499	3,740822506
X917	6,628455558	X378	4,39843773	X359	4,006998315
X410	4,766380121	X398	3,561190917	X61	3,975698256
X309	4,77683088	X563	7,617578564	X77	3,043480519
X728	2,02497137	X794	1,894753296	X197	3,085823206
X711	2,920723278	X757	2,639831737		

Appendix E

Feature importance for one fold in the replication performed

Feature	%IncMSE
MS.HL	63.030802095337
APC_wt	51.6808498904511
BRAF_cn	-0.356927579747502
BRCA2_cn	0.144166288185375
CCND3_cn	0.112519326209813
CDK4_cn	15.4450345132938
CDKN2A_cn	134.559803938746
CDKN2a.p14._wt	154.198440218973
CYLD_wt	3.01170183363915
EP300_cn	15.1159710814059
EZH2_wt	-18.2199215052158
FAM123B_cn	12.2189309382264
FGFR2_cn	30.5273389366171
FLCN_wt	3.06406316035099
GNAS_wt	23.4765747385554
IDH1_wt	16.7705293751201
JAK2_cn	-2.8236868117502
KDM6A_wt	18.4310850418543
KIT_cn	13.6368894063793
MAP2K4_wt	11.0823719021441
MET_cn	18.5785936682867
MLLT3_cn	-13.6973756182997
MSH6_wt	-7.80250627512262
MYCL1_cn	6.89120795047741
NF1_cn	-2.53422075484848
NOTCH1_wt	36.5730415370446
PDGFRA_cn	-16.4232965164592
PIK3R1_wt	25.9892237549616
PTEN_cn	53.8248669431963
RUNX1_wt	10.7548728340524
SMAD4_cn	1.53439590654189
SMARCB1_cn	-6.87226915527052
STK11_wt	52.3333522918816
TP53_wt	60.9882620382466
TSC1_cn	12.0933316735095
WT1_cn	6.65157991583468
PubchemFP2	-1.77972333038301
PubchemFP10	2.49556635136749
PubchemFP13	1.86596520618869
PubchemFP16	5.08906457968508
PubchemFP19	7.2573600663377
PubchemFP22	2.65606641304926
PubchemFP25	3.52451893668916
PubchemFP34	5.05741216626264
PubchemFP43	1.50330885675088
PubchemFP93	3.88433359113152
PubchemFP117	-0.686177687322741
PubchemFP130	0.352518730536206
PubchemFP144	3.07076453164156
PubchemFP147	-3.3945566625045
PubchemFP150	5.05684588462767
PubchemFP155	5.97871491011471
PubchemFP159	2.25564349555199
PubchemFP167	1.60074228623241
PubchemFP180	7.61527716222515
PubchemFP183	3.89790805057656
PubchemFP186	3.91759982639011
PubchemFP189	1.08434370664216
PubchemFP192	5.48463086995993
PubchemFP195	3.11477408860776
PubchemFP206	5.98406765315526
PubchemFP218	1.65651656254834
AKT2_cn	15.5944570109132
APC_cn	10.4709345146404
BRCA1_wt	26.1892314416616
CCND1_cn	31.7505936438385
CDH1_wt	-20.389234925535
CDK6_cn	-6.97465697925511
CDKN2C_wt	3.70554542329518
CTNNB1_wt	19.3315323455614
EGFR_wt	5.24854272995059
ERBB2_wt	7.31434498983231
EZH2_cn	0.0293511992847435
FBXW7_wt	32.039180555225
FGFR3_wt	11.586452237742
FLT3_wt	5.222034146487
GNAS_cn	10.9797220575091
IDH1_cn	7.83003498191086
KDM5C_wt	7.3695414635145
KDM6A_cn	-1.73537172805421
KRAS_wt	102.893368963218
MAP2K4_cn	4.02583849026277
MLH1_wt	56.8877905521721
MSH2_wt	-6.50901792897395
MSH6_cn	7.38980159212067
MYCN_cn	35.6088423704808
NF2_wt	41.6774208038113
NRAS_wt	52.8589213170972
PIK3CA_wt	43.8555270689856
PIK3R1_cn	15.3816181805634
RB1_wt	33.6458614785392
SETD2_wt	-14.1359033254016
SMARCA4_wt	14.4325800708993
SMO_cn	-2.39666286216357
STK11_cn	4.26985928256955
TP53_cn	2.26316427755261
VHL_wt	14.8040232426642
PubchemFP0	5.39288697709807
PubchemFP6	8.84413888340566
PubchemFP11	1.45294688616572
PubchemFP14	3.4672351672435
PubchemFP17	2.19146711852432
PubchemFP20	6.94433117218408
PubchemFP23	9.4182492005021
PubchemFP30	1.44870839033914
PubchemFP37	5.45253293741407
PubchemFP44	2.01560458319998
PubchemFP115	2.83712082578723
PubchemFP118	-1.81090964529685
PubchemFP132	-1.2187245040104
PubchemFP145	6.35720966361323
PubchemFP148	6.94921425063049
PubchemFP152	5.43959180748051
PubchemFP156	5.80646512950928
PubchemFP160	2.90367920340759
PubchemFP178	12.7933050967614
PubchemFP181	7.03649332722284
PubchemFP184	5.4879003455738
PubchemFP187	15.152462944287
PubchemFP190	-5.93513622311873
PubchemFP193	7.23279900664838
PubchemFP199	4.37744797771988
PubchemFP213	1.14174431201675
PubchemFP219	2.15786895338216
ALK_wt	1.7995286469141
BRAF_wt	133.609599865903
BRCA2_wt	3.6351528948809
CCND2_cn	27.1913512687925
CDH1_cn	10.4039287699489
CDKN2A_wt	122.168537937684
CDKN2C_cn	14.09721518872
CTNNB1_cn	14.8825579126525
EGFR_cn	-13.6234880523058
ERBB2_cn	56.2161801617159
FAM123B_wt	14.9641852744136
FBXW7_cn	3.44306779276597
FGFR3_cn	10.5450103493052
FLT3_cn	-0.918992968050054
HRAS_wt	-24.0405580080076
JAK2_wt	10.7993478309974
KDM5C_cn	-0.93844790088401
KDR_cn	8.90065886863179
KRAS_cn	-17.9162712942346
MDM2_cn	33.7320988208828
MLH1_cn	7.07995632988188
MSH2_cn	3.6961685057127
MYC_cn	29.5648063247257
NF1_wt	26.874611205512
NF2_cn	12.1074031513538
NRAS_cn	21.8701033091377
PIK3CA_cn	21.6705918990256
PTEN_wt	61.9079055825935
RB1_cn	2.02301926322838
SMAD4_wt	111.197026628522
SMARCA4_cn	-4.29663440852717
SOCS1_cn	-0.255989690729639
TET2_wt	39.482823904137
TSC1_wt	22.7790708292619
VHL_cn	16.3108288684174
PubchemFP1	1.5446031270372
PubchemFP9	0.868401836969371
PubchemFP12	2.50721861126388
PubchemFP15	4.55177127958497
PubchemFP18	3.24370397568824
PubchemFP21	1.25996663183254
PubchemFP24	6.95477874511905
PubchemFP33	4.40305791627611
PubchemFP38	5.40059446571981
PubchemFP46	4.77065189765585
PubchemFP116	2.818701183711
PubchemFP129	0.831293692484912
PubchemFP143	4.34649972961588
PubchemFP146	6.48600224058122
PubchemFP149	5.65035199268905
PubchemFP153	4.09270614497437
PubchemFP157	9.01332197364307
PubchemFP164	1.79048951623869
PubchemFP179	2.82517545386095
PubchemFP182	6.08408845074705
PubchemFP185	2.59084758557437
PubchemFP188	12.5749090751241
PubchemFP191	3.46772106624199
PubchemFP194	6.59201497003662
PubchemFP200	3.01007733515858
PubchemFP214	-5.245803004306
PubchemFP227	-6.43754828402522

PubchemFP228	1.04104769696486	PubchemFP232	-5.407332655547834	PubchemFP233	-4.480201603201870
PubchemFP241	-0.309631920489828	PubchemFP246	-0.66358207196537	PubchemFP247	0.613054065054675
PubchemFP248	1.55389450337808	PubchemFP252	0.468147748180114	PubchemFP255	2.04262338381516
PubchemFP256	4.75849751635205	PubchemFP257	4.28958192856559	PubchemFP258	8.0276761466956
PubchemFP259	3.79347625438562	PubchemFP260	4.07398400367559	PubchemFP261	5.33490916421802
PubchemFP262	1.63659442904341	PubchemFP274	8.06261425076962	PubchemFP276	7.76047367282502
PubchemFP283	1.60905326675799	PubchemFP284	3.01421091124859	PubchemFP285	2.64768597297803
PubchemFP286	3.75600175647762	PubchemFP287	10.4457054833001	PubchemFP293	2.37532521308849
PubchemFP294	6.05999138561047	PubchemFP297	3.03728065282046	PubchemFP298	5.45869320730233
PubchemFP299	5.94754786491798	PubchemFP300	5.23054548924839	PubchemFP301	10.7432827788657
PubchemFP305	3.29972775503794	PubchemFP308	7.9083341983648	PubchemFP314	1.88542563087457
PubchemFP327	2.78534561737192	PubchemFP328	2.68070980858481	PubchemFP330	2.8442205367358
PubchemFP332	1.77739106796962	PubchemFP333	2.10856848150883	PubchemFP334	5.69846155054877
PubchemFP335	2.93716647790619	PubchemFP336	5.18809360329488	PubchemFP337	21.1251791965691
PubchemFP338	4.81422972301471	PubchemFP339	5.48105195600646	PubchemFP340	4.16638593761856
PubchemFP341	6.68302203020198	PubchemFP342	4.99261999516781	PubchemFP344	2.14634715556342
PubchemFP345	5.73822594267683	PubchemFP346	8.52441184206974	PubchemFP347	2.64672171074501
PubchemFP349	6.80400799706548	PubchemFP350	4.44717872920241	PubchemFP351	3.38293454002848
PubchemFP352	2.95944392670927	PubchemFP353	4.65468188773278	PubchemFP355	1.8498267391067
PubchemFP356	4.13989013800825	PubchemFP357	5.7046958832579	PubchemFP358	3.89575617906114
PubchemFP359	4.89902647004964	PubchemFP360	0.185821847949988	PubchemFP362	6.86403135012526
PubchemFP363	4.18841439903335	PubchemFP364	5.3571718471372	PubchemFP365	4.55400386698942
PubchemFP366	6.30565784696109	PubchemFP367	3.08541076614299	PubchemFP368	2.40238208503982
PubchemFP370	2.922051951852	PubchemFP371	1.43875205889835	PubchemFP372	5.10146453680511
PubchemFP373	5.0618103882342	PubchemFP374	9.34453525209538	PubchemFP375	6.56995784404531
PubchemFP376	2.94850225055982	PubchemFP377	9.02423513632515	PubchemFP378	4.03130837456457
PubchemFP379	2.47902998255392	PubchemFP380	1.90460914431438	PubchemFP381	3.95031032453805
PubchemFP382	2.39693902860443	PubchemFP383	2.52443220147329	PubchemFP384	1.67475596436154
PubchemFP385	4.07565714256484	PubchemFP386	8.37096375291921	PubchemFP387	3.82817300792413
PubchemFP388	3.5096507318583	PubchemFP389	3.91221038429244	PubchemFP390	3.32322129843472
PubchemFP391	5.62159289206657	PubchemFP392	6.07640028177484	PubchemFP393	5.38158778472045
PubchemFP394	3.18109328906227	PubchemFP395	11.3325386869067	PubchemFP396	2.91467294561337
PubchemFP397	3.20551936854935	PubchemFP398	3.97116879407602	PubchemFP399	3.13388110685478
PubchemFP400	3.93356858037794	PubchemFP403	4.67727361133478	PubchemFP404	1.29698602796767
PubchemFP405	5.1275429224587	PubchemFP406	5.76126229081497	PubchemFP407	1.63067743756927
PubchemFP408	3.59526742419134	PubchemFP409	1.81984392142365	PubchemFP411	1.94010970209228
PubchemFP412	3.8459382770605	PubchemFP413	2.68491324709903	PubchemFP414	1.36712403274931
PubchemFP416	1.73551193673471	PubchemFP417	2.88996909145362	PubchemFP418	3.34797915473963
PubchemFP419	4.31854535910808	PubchemFP420	5.87355239964331	PubchemFP421	3.20878529544321
PubchemFP422	4.65259373476886	PubchemFP423	2.69567671270907	PubchemFP425	1.94409055979521
PubchemFP427	2.95093349051344	PubchemFP428	1.88098220171213	PubchemFP429	4.0483558367319
PubchemFP430	4.56822504711882	PubchemFP431	5.37273369196822	PubchemFP432	6.77467026410277
PubchemFP434	6.12099386042609	PubchemFP435	4.65398944866572	PubchemFP436	1.60713451827923
PubchemFP437	6.70440779707276	PubchemFP438	4.17570827001739	PubchemFP439	9.33022128672683
PubchemFP440	3.84564109266531	PubchemFP441	1.57503592561199	PubchemFP442	1.89068086730718
PubchemFP443	6.77712137438782	PubchemFP445	6.12819279620185	PubchemFP446	7.14796991508087
PubchemFP447	6.00680090594806	PubchemFP448	0.372792695608645	PubchemFP449	5.50516775015453
PubchemFP450	5.21918842891525	PubchemFP451	4.47471208103286	PubchemFP452	5.18273030039527
PubchemFP453	4.19749746532141	PubchemFP454	2.01692454752922	PubchemFP456	1.12735564479396
PubchemFP457	1.60718972812291	PubchemFP458	3.51984687277312	PubchemFP459	3.37130612341431
PubchemFP460	2.85266579809528	PubchemFP461	3.33177583990828	PubchemFP462	3.18599434048464
PubchemFP464	3.60750159194424	PubchemFP465	3.90677506855977	PubchemFP466	5.52012303862084
PubchemFP467	3.39460650696399	PubchemFP469	-5.26721868140873	PubchemFP470	1.6180970561359
PubchemFP471	5.42360691101173	PubchemFP472	4.67507549600245	PubchemFP473	-0.814209982952209
PubchemFP474	0.401329694172159	PubchemFP475	-1.56324253977545	PubchemFP476	4.55216732134435
PubchemFP477	5.14785502585983	PubchemFP480	3.24333623997205	PubchemFP481	1.30876010660605
PubchemFP482	3.07146429821919	PubchemFP483	1.96559736229943	PubchemFP484	2.99495862344323
PubchemFP485	5.93458072848073	PubchemFP486	1.47225521340665	PubchemFP487	6.03708878810751
PubchemFP489	3.92252978506795	PubchemFP490	-0.9280521714391	PubchemFP493	6.63140863030357
PubchemFP494	3.98229658846159	PubchemFP495	3.28696532373533	PubchemFP497	3.46772335273656
PubchemFP498	7.70609625000351	PubchemFP499	6.97276370633235	PubchemFP500	3.98479696672869
PubchemFP501	4.50592487403344	PubchemFP504	3.78859803000591	PubchemFP505	1.60981016306985
PubchemFP506	4.59716260151855	PubchemFP507	2.99515867852789	PubchemFP508	3.55808009853669
PubchemFP509	2.55208116094845	PubchemFP514	10.9918308545105	PubchemFP515	4.55128831261286
PubchemFP516	2.51288511619888	PubchemFP517	2.77771869519761	PubchemFP518	2.26737481746071
PubchemFP519	4.6138733903676	PubchemFP520	1.94116788257729	PubchemFP521	1.58288340364081
PubchemFP523	3.56768282254476	PubchemFP524	1.10694547700385	PubchemFP530	3.30851941482973
PubchemFP531	3.2378356836697	PubchemFP533	3.31314008662504	PubchemFP534	1.95351664451381
PubchemFP535	8.83534787145628	PubchemFP536	6.94372564053958	PubchemFP537	3.39601336048591
PubchemFP538	4.80298461999938	PubchemFP539	-1.31847614534175	PubchemFP540	6.93140954412761
PubchemFP541	4.47114643376621	PubchemFP543	5.82551019491804	PubchemFP544	4.72603200421132
PubchemFP545	2.99474793858952	PubchemFP547	6.16812717225098	PubchemFP548	11.3010203611482
PubchemFP549	2.7857825431307	PubchemFP550	4.87871044467945	PubchemFP551	3.30743855365824
PubchemFP552	0.976718408651143	PubchemFP553	6.4737505907108	PubchemFP554	2.55004179417825
PubchemFP555	2.68378730604056	PubchemFP556	-0.0686857737248161	PubchemFP558	3.08641772865379
PubchemFP560	6.28554178091023	PubchemFP563	10.6732220040996	PubchemFP564	-0.445371896518721
PubchemFP565	4.53087247311992	PubchemFP566	7.17554139962095	PubchemFP567	5.00127608787197
PubchemFP568	5.4455399003224	PubchemFP569	4.25020973642612	PubchemFP570	2.30777293202477
PubchemFP572	4.85725342307502	PubchemFP573	4.47504018434535	PubchemFP574	9.76625252156855
PubchemFP575	4.02880713759012	PubchemFP577	6.29298535672695	PubchemFP578	2.5749380551831
PubchemFP579	8.26442243570094	PubchemFP580	4.66391204871644	PubchemFP581	27.9696333388053
PubchemFP582	1.84037036112381	PubchemFP583	2.39764652202969	PubchemFP584	0.969477165992181
PubchemFP585	3.34929471103096	PubchemFP586	4.92269702497268	PubchemFP588	2.03007766930869
PubchemFP589	2.2732471889675	PubchemFP591	5.09692096324374	PubchemFP592	2.75131363655324
PubchemFP593	4.54350841374802	PubchemFP594	2.79685799288655	PubchemFP595	2.08518926057958
PubchemFP596	2.60368011892957	PubchemFP597	6.14590987361902	PubchemFP598	4.57932977987381
PubchemFP599	3.68396556548328	PubchemFP600	5.2884064814872	PubchemFP601	4.68228623986897
PubchemFP602	7.39870667345028	PubchemFP603	2.07463687344364	PubchemFP604	10.8641684868701
PubchemFP605	3.96634076715873	PubchemFP606	2.49122441987923	PubchemFP607	2.75960044424834
PubchemFP608	2.29283584713639	PubchemFP609	4.32009683062746	PubchemFP610	2.13691375432355
PubchemFP611	4.2716988873397	PubchemFP612	4.67553129458909	PubchemFP613	4.22698952030245
PubchemFP614	4.00395937732994	PubchemFP615	5.65792372189169	PubchemFP616	5.11604468772715
PubchemFP618	1.69037843891179	PubchemFP619	3.82511272781075	PubchemFP620	2.52987672527921

PubchemFP621	3.17497870167321	PubchemFP622	-3.02296078745084	PubchemFP623	5.06394296582221
PubchemFP624	4.71623850307325	PubchemFP625	5.55451514734	PubchemFP626	4.39100495412665
PubchemFP628	2.29632140490542	PubchemFP629	4.4223016640676	PubchemFP630	2.40844067796125
PubchemFP632	4.08708178732379	PubchemFP633	4.29359596074767	PubchemFP634	1.42667250315176
PubchemFP636	3.43649989898118	PubchemFP637	5.52996968332022	PubchemFP638	5.09714454423596
PubchemFP639	2.69481924162362	PubchemFP640	1.88757879795377	PubchemFP641	4.53769958143171
PubchemFP642	27.0601167724226	PubchemFP644	2.84113840804731	PubchemFP645	9.86212787862481
PubchemFP646	2.6321840370175	PubchemFP647	2.07378140282095	PubchemFP648	1.75248364303548
PubchemFP650	4.42575979416326	PubchemFP651	4.20100687021814	PubchemFP652	4.47385012169164
PubchemFP653	2.82027885072289	PubchemFP654	4.37603537236683	PubchemFP655	6.44163779513853
PubchemFP656	4.03493612939661	PubchemFP657	2.92315819106799	PubchemFP658	7.22433660905773
PubchemFP659	7.17280895665109	PubchemFP660	0.958333421722222	PubchemFP661	11.458658877799
PubchemFP662	1.72082057348523	PubchemFP664	1.6945816021818	PubchemFP665	5.45115928656079
PubchemFP666	3.89269094875753	PubchemFP668	1.30863217728908	PubchemFP669	2.0895920122979
PubchemFP670	1.93228403229391	PubchemFP671	6.7302239130789	PubchemFP673	7.09929022582835
PubchemFP674	3.9119424096432	PubchemFP675	2.15500675496966	PubchemFP676	4.82736279895433
PubchemFP677	2.5293326431561	PubchemFP678	0.197702059495728	PubchemFP679	1.20592524141585
PubchemFP680	4.0622691426491	PubchemFP681	6.7419978822189	PubchemFP682	8.51689126994464
PubchemFP683	4.14336952712752	PubchemFP684	5.95082235402873	PubchemFP685	5.82310763557059
PubchemFP686	11.9195790396474	PubchemFP687	8.74591071291774	PubchemFP688	7.95761425981889
PubchemFP689	4.50133437206196	PubchemFP690	4.44327896832754	PubchemFP691	7.18718209149912
PubchemFP692	7.35866348098259	PubchemFP693	2.94815080252367	PubchemFP694	1.49549911204744
PubchemFP695	5.90317108706607	PubchemFP696	4.43363777984911	PubchemFP697	12.064287358764
PubchemFP698	7.75385074638105	PubchemFP699	4.84872784625022	PubchemFP700	3.83975380672992
PubchemFP701	4.84094504516875	PubchemFP702	6.4941292407227	PubchemFP703	5.52189139152706
PubchemFP704	6.25731917230476	PubchemFP705	5.20363699174505	PubchemFP706	-3.84713439720219
PubchemFP707	5.63051099370526	PubchemFP708	2.02458060596262	PubchemFP709	3.72399950866053
PubchemFP710	4.89398859268389	PubchemFP711	2.31212735130673	PubchemFP712	6.51287254474513
PubchemFP713	7.82788633718858	PubchemFP714	4.78624977887842	PubchemFP715	2.18320397418448
PubchemFP716	7.75276579735395	PubchemFP717	3.23486059966949	PubchemFP719	2.78037432362557
PubchemFP721	5.79794679493236	PubchemFP722	2.70014370866021	PubchemFP725	1.91898883345123
PubchemFP728	8.96020904281336	PubchemFP729	2.26247791222309	PubchemFP733	2.13005127847135
PubchemFP734	7.26393265336201	PubchemFP735	5.63684177371791	PubchemFP736	2.68622466683074
PubchemFP737	4.7042622371881	PubchemFP738	4.62159281242579	PubchemFP740	2.82308005477814
PubchemFP742	3.72697750998243	PubchemFP743	2.12953045525639	PubchemFP745	0.0965815318655915
PubchemFP746	4.47757572788864	PubchemFP747	3.10203560533094	PubchemFP749	5.79671436381013
PubchemFP750	7.36607070220743	PubchemFP751	2.52895272812964	PubchemFP752	2.88374365578634
PubchemFP755	7.61766951028283	PubchemFP756	3.18285065818063	PubchemFP757	3.97615585482496
PubchemFP758	5.83777271077832	PubchemFP759	5.03059456806628	PubchemFP761	3.44012266495051
PubchemFP762	-1.35063748856266	PubchemFP763	4.57950724350519	PubchemFP764	1.29667013673279
PubchemFP766	2.2768855199181	PubchemFP767	2.59930089137576	PubchemFP770	3.70987345977093
PubchemFP771	6.23106196522795	PubchemFP772	2.23263494755983	PubchemFP776	6.32974621974487
PubchemFP777	3.78456347039323	PubchemFP778	2.67462656524088	PubchemFP779	8.34264400242892
PubchemFP780	3.63263890476523	PubchemFP782	3.76116872125897	PubchemFP784	5.36957295907781
PubchemFP785	2.27504460619906	PubchemFP788	0.338941399589503	PubchemFP791	9.15982669586266
PubchemFP792	2.26495409767388	PubchemFP796	1.78447252756069	PubchemFP797	4.9497341184137
PubchemFP798	6.13544148535155	PubchemFP799	2.33774904519446	PubchemFP800	4.23295693769458
PubchemFP801	4.48256769611401	PubchemFP803	2.81362211500564	PubchemFP805	4.2996488393107
PubchemFP806	1.51978947975468	PubchemFP808	1.50328685991531	PubchemFP809	3.02882666337605
PubchemFP810	1.7826981139546	PubchemFP812	7.1862488055385	PubchemFP813	7.68098343301358
PubchemFP814	2.90107068454887	PubchemFP815	2.01944324185055	PubchemFP818	6.41825368458259
PubchemFP819	5.99202556570906	PubchemFP820	4.04767113162199	PubchemFP821	5.89612814381938
PubchemFP822	4.31272861342466	PubchemFP824	3.68546224899147	PubchemFP825	-4.30129695367445
PubchemFP826	5.32174039421538	PubchemFP827	2.207521843622	PubchemFP829	1.62528831490701
PubchemFP830	2.91065006406725	PubchemFP833	3.62178044505095	PubchemFP834	4.17354658912499
PubchemFP835	2.15049083106929	PubchemFP839	1.33275964914727	PubchemFP840	-4.32941994444933
PubchemFP860	-3.84860754898041	PubchemFP861	-4.87338272744492	nAcid	3.49490130539136
apol	8.48102679142272	naAromAtom	13.8925495441836	nAromBond	12.6210249219444
nAtom	7.12939176939597	nHeavyAtom	4.99173586322448	nH	15.9418528358722
nB	9.27965047936655	nC	12.5529729732747	nN	7.05734607567905
nO	9.95508199545984	nS	6.6745996736291	nP	1.0898215540106
nF	10.730670521147	nCl	8.88726961794345	nBr	1.63459614396592
nI	4.51569500115588	ATSc1	13.4689406262689	ATSc2	18.2322812492603
ATSc3	30.4195755388618	ATSc4	22.8056178174134	ATSc5	23.5315343023311
ATSm1	12.8377251815208	ATSm2	8.06156518811558	ATSm3	6.0189999387248
ATSm4	10.4470480358729	ATSm5	10.566960878491	ATSp1	9.04873423973322
ATSp2	7.8206102007326	ATSp3	10.2001215560857	ATSp4	10.3443349627243
ATSp5	9.87404272910894	nBase	11.5100191470814	nBonds	6.29532067260421
nBonds2	10.769935480839	nBondsS	11.6043143609767	nBondsS2	10.1707400194455
nBondsS3	11.9643891195926	nBondsD	13.0307317692353	nBondsD2	15.0423083722397
nBondsT	3.0191766944979	bpol	7.25398232221872	C1SP1	3.84094222664992
C2SP1	2.95845540924336	C1SP2	10.7701568308379	C2SP2	19.0279216432238
C3SP2	21.3609908441646	C1SP3	13.9613625059444	C2SP3	10.8476864277824
C3SP3	9.03812138075901	C4SP3	4.78372134664915	SCH.3	2.91348509221845
SCH.4	7.73312698567442	SCH.5	13.0543283620492	SCH.6	19.4752096154005
SCH.7	18.0860996023325	VCH.3	3.68491877170265	VCH.4	8.20131055348199
VCH.5	15.4296040148552	VCH.6	15.4113735217743	VCH.7	24.6603290395259
SC.3	13.1940917578723	SC.4	14.9027251767708	SC.5	7.7793387086907
SC.6	12.5212770903705	VC.3	13.1084296457726	VC.4	24.4059536856909
VC.5	15.3416157389446	VC.6	12.0148644181644	SPC.4	9.61194894724826
SPC.5	12.4794186073521	SPC.6	10.1055271683703	VPC.4	12.7752221309617
VPC.5	11.8880440746878	VPC.6	12.3746726427276	ECCEN	12.8950563960889
fragC	15.8362958824824	nHBAcc	7.87080518864251	nHBAcc2	9.55755183446104
nHBAcc3	18.2863585121702	nHBAcc_Lipinski	6.43537887525071	nHBDon	30.5883830110861
nHBDon_Lipinski	11.4079213169022	nAtomLC	12.9082762665911	nAtomP	14.9174093692689
nAtomLAC	11.7173162580729	MLogP	24.3053918966804	McGowan_Volume	10.6627209176955
MDEC.11	9.48250616647548	MDEC.12	11.9277826502012	MDEC.13	14.660641146342
MDEC.14	11.0529690892909	MDEC.22	23.5675769414648	MDEC.23	13.2928795754932
MDEC.24	16.964993807552	MDEC.33	14.1633464639132	MDEC.34	17.6183940316126
MDEC.44	3.93356944754529	MDEO.11	19.7373362333804	MDEO.12	40.801185683445
MDEO.22	4.34403408294465	MDEN.11	9.0106577902956	MDEN.12	10.2103834721596
MDEN.13	28.5854186527228	MDEN.22	14.9454125602534	MDEN.23	12.2460571986371
MDEN.33	10.184012824124	MLFER_A	14.1597054638882	MLFER_BH	13.8323440432494

MLFER_BO	19.7038587126037	MLFER_S	11.7461094465931	MLFER_E	17.3803523769754
MLFER_L	25.4229283688413	PetitjeanNumber	15.9210997747369	nRing	6.01002682580881
n3Ring	1.54277734155562	n4Ring	1.2968866445264	n5Ring	9.72420155380451
n6Ring	13.3742282437302	n7Ring	0.926358329656961	n8Ring	-4.91969951265647
n9Ring	0.977539993342432	n10Ring	0.766230159511356	nG12Ring	2.58521494459324
nFRing	8.51980532310474	nF6Ring	-0.138635046570875	nF8Ring	2.19925034660949
nF9Ring	7.41628130436968	nF10Ring	11.003879064936	nF11Ring	2.26180826388436
nF12Ring	5.42505362847208	nFG12Ring	8.71912337688775	nTRing	10.1611165773085
nT4Ring	0.941460825224625	nT5Ring	9.78110980250178	nT6Ring	13.9143029733709
nT7Ring	1.01217580346026	nT8Ring	1.31055762509429	nT9Ring	8.46370117148573
nT10Ring	9.88445788351116	nT11Ring	3.12038276872752	nT12Ring	4.60550749534969
nTG12Ring	7.94720887311358	nRotB	14.3531351555421	TopoPSA	18.7343576012179
VAdjMat	5.18745526813097	MW	11.6759680577346	WTPT.1	6.94995896695379
WTPT.2	15.0451696161186	WTPT.3	12.5384014051668	WTPT.4	17.9765807321768
WTPT.5	15.2343458543523	WPATH	18.493999907808	WPOL	9.27140414065025
Zagreb	7.66041017301296				

Bibliography

[1] B. W. Stewart and C. P. Wild, *World Cancer Report 2014*, English. Lyon: International Agency for Research on Cancer/World Health Organization, 2014, OCLC: 908606220, ISBN: 978-92-832-0443-5.

[2] B. A. Kohler, R. L. Sherman, N. Howlader, A. Jemal, A. B. Ryerson, K. A. Henry, F. P. Boscoe, K. A. Cronin, A. Lake, A.-M. Noone, S. J. Henley, C. R. Eheman, R. N. Anderson, and L. Penberthy, "Annual Report to the Nation on the Status of Cancer, 1975-2011, Featuring Incidence of Breast Cancer Subtypes by Race/Ethnicity, Poverty, and State", eng, *Journal of the National Cancer Institute*, vol. 107, no. 6, djv048, Jun. 2015, ISSN: 1460-2105. DOI: 10.1093/jnci/djv048.

[3] G. J. Kelloff and C. C. Sigman, "Cancer biomarkers: Selecting the right drug for the right patient", *Nature reviews Drug discovery*, vol. 11, no. 3, pp. 201–214, 2012. [Online]. Available: http://www.nature.com/nrd/journal/v11/n3/abs/nrd3651.html.

[4] L. Moja, L. Tagliabue, S. Balduzzi, E. Parmelli, V. Pistotti, V. Guarneri, and R. D'Amico, "Trastuzumab containing regimens for early breast cancer", en, in *Cochrane Database of Systematic Reviews*, DOI: 10.1002/14651858.CD006243.pub2, John Wiley & Sons, Ltd, Apr. 2012. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD006243.pub2/abstract.

[5] M. P. Menden, F. Iorio, M. Garnett, U. McDermott, C. H. Benes, P. J. Ballester, and J. Saez-Rodriguez, "Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties", en, *PLoS ONE*, vol. 8, no. 4, G. P. S. Raghava, Ed., e61318, Apr. 2013, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0061318. [Online]. Available: http://dx.plos.org/10.1371/journal.pone.0061318.

[6] J. T. C. M. Ladeiras, "Previsão de respostas a tratamentos de linhas celulares cancerígenas", 2015. [Online]. Available: https://repositorio-aberto.up.pt/handle/10216/83478.

[7] S. Papillon-Cavanagh, N. De Jay, N. Hachem, C. Olsen, G. Bontempi, H. J. W. L. Aerts, J. Quackenbush, and B. Haibe-Kains, "Comparison and validation of genomic predictors for anticancer drug sensitivity", *Journal of the American Medical Informatics Association*, vol. 20, no. 4, pp. 597–602, Jul. 2013, ISSN: 1067-5027. DOI: 10.1136/amiajnl-2012-001442. [Online]. Available: https://academic.oup.com/jamia/article/20/4/597/2909333/Comparison-and-validation-of-genomic-predictors.

[8] National Cancer Institute, *What Is Cancer?*, cgvArticle. [Online]. Available: https://www.cancer.gov/about-cancer/understanding/what-is-cancer.

[9] W. H. Organization, *Cancer Fact Sheet*. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs297/en/.

[10] K. Polyak, M. Shipitsin, L. Campbell-Marrotta, N. Bloushtain-Qimron, and S. Y. Park, "Breast tumor heterogeneity: Causes and consequences", *Breast Cancer Research*, vol. 11, S18, Jun. 2009, ISSN: 1465-542X. DOI: 10.1186/bcr2279. [Online]. Available: https://doi.org/10.1186/bcr2279.

[11] S. V. Sharma, D. A. Haber, and J. Settleman, "Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents", *Nature Reviews Cancer*, vol. 10, no. 4, pp. 241–253, Apr. 2010, ISSN: 1474-175X, 1474-1768. DOI: 10.1038/nrc2820. [Online]. Available: http://www.nature.com/doifinder/10.1038/nrc2820.

[12] K. Strimbu and J. A. Tavel, "What are biomarkers?:" en, *Current Opinion in HIV and AIDS*, vol. 5, no. 6, pp. 463–466, Nov. 2010, ISSN: 1746-630X. DOI: 10.1097/COH.0b013e32833ed177. [Online]. Available: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=01222929-201011000-00003.

[13] Arshad Chaudry, "Cell Culture", *The Science Creative Quarterly*, Aug. 2004. [Online]. Available: http://www.scq.ubc.ca/cell-culture/.

[14] J. R. Masters, "Human cancer cell lines: Fact and fantasy", eng, *Nature Reviews. Molecular Cell Biology*, vol. 1, no. 3, pp. 233–236, Dec. 2000, ISSN: 1471-0072. DOI: 10.1038/35043102.

[15] J. L. Sebaugh, "Guidelines for accurate EC50/IC50 estimation", eng, *Pharmaceutical Statistics*, vol. 10, no. 2, pp. 128–134, Apr. 2011, ISSN: 1539-1612. DOI: 10.1002/pst.426.

[16] J. S. Soothill, R. Ward, and A. J. Girling, "The IC50: An exactly defined measure of antibiotic sensitivity", *Journal of Antimicrobial Chemotherapy*, vol. 29, no. 2, pp. 137–139, Feb. 1992, ISSN: 0305-7453. DOI: 10.1093/jac/29.2.137. [Online]. Available: https://academic.oup.com/jac/article/29/2/137/683730/The-IC50-an-exactly-defined-measure-of-antibiotic.

[17] The Human Genome Project, *What is a genome?* [Online]. Available: https://ghr.nlm.nih.gov/primer/hgp/genome.

[18] J. R. Pollack, C. M. Perou, A. A. Alizadeh, M. B. Eisen, A. Pergamenschikov, C. F. Williams, S. S. Jeffrey, D. Botstein, and P. O. Brown, "Genome-wide analysis of DNA copy-number changes using cDNA microarrays", eng, *Nature Genetics*, vol. 23, no. 1, pp. 41–46, Sep. 1999, ISSN: 1061-4036. DOI: 10.1038/12640.

[19] M. Zarrei, J. R. MacDonald, D. Merico, and S. W. Scherer, "A copy number variation map of the human genome", en, *Nature Reviews Genetics*, vol. 16, no. 3, pp. 172–183, Mar. 2015, ISSN: 1471-0056. DOI: 10.1038/nrg3871. [Online]. Available: http://www.nature.com/nrg/journal/v16/n3/full/nrg3871.html.

[20] National Cancer Institute, *Definition of phenotype*, nciAppModulePage. [Online]. Available: https://www.cancer.gov/publications/dictionaries/genetics-dictionary.

[21] L. A. Loeb, "A Mutator Phenotype in Cancer", en, *Cancer Research*, vol. 61, no. 8, pp. 3230–3239, Apr. 2001, ISSN: 0008-5472, 1538-7445. [Online]. Available: http://cancerres.aacrjournals.org/content/61/8/3230.

[22] U. Shokal and P. C. Sharma, "Implication of microsatellite instability in human gastric cancers", *The Indian Journal of Medical Research*, vol. 135, no. 5, pp. 599–613, May 2012, ISSN: 0971-5916. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3401689/.

[23] F. K. Brown, "Chapter 35 - Chemoinformatics: What is it and How does it Impact Drug Discovery.", in *Annual Reports in Medicinal Chemistry*, J. A. Bristol, Ed., vol. 33, DOI: 10.1016/S0065-7743(08)61100-8, Academic Press, Jan. 1998, pp. 375–384. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0065774308611008.

[24] OECD, *Introduction to (Quantitative) Structure Activity Relationships*. [Online]. Available: http://www.oecd.org/env/ehs/risk-assessment/introductiontoquantitativestruc htm.

[25] R. Todeschini, V. Consonni, and R. Mannhold, *Molecular Descriptors for Chemoinformatics: Volume I: Alphabetical Listing / Volume II: Appendices, References*, en. Wiley-VCH, Sep. 2009, Google-Books-ID: DcrwAAAAMAAJ, ISBN: 978-3-527-31852-0.

[26] A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. I. Gushurst, D. L. Grier, B. A. Leland, and J. Laufer, "Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited", *Journal of Chemical Information and Computer Sciences*, vol. 32, no. 3, pp. 244–255, May 1992, ISSN: 0095-2338. DOI: 10.1021/ci00007a012. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ci00007a012.

[27] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open Babel: An open chemical toolbox", *Journal of Cheminformatics*, vol. 3, p. 33, Oct. 2011, ISSN: 1758-2946. DOI: 10.1186/1758-2946-3-33. [Online]. Available: https://doi.org/10.1186/1758-2946-3-33.

[28] C. W. Yap, "PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints", eng, *Journal of Computational Chemistry*, vol. 32, no. 7, pp. 1466–1474, May 2011, ISSN: 1096-987X. DOI: 10.1002/jcc.21707.

[29] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen, "The Chemistry Development Kit (CDK): An open-source Java library for Chemo- and Bioinformatics", eng, *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 2, pp. 493–500, Apr. 2003, ISSN: 0095-2338. DOI: 10.1021/ci025584y.

[30] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, en. Elsevier, 2012, Google-Books-ID: dfs2kgEACAAJ, ISBN: 978-93-80931-91-3.

[31] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining", in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, Citeseer, 2000, pp. 29–39.

[32] IBM, *IBM Knowledge Center - CRISP-DM Help Overview*, en-US. [Online]. Available: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.crispdm.help/crisp_overview.htm.

[33] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, ser. Springer Texts in Statistics. New York, NY: Springer New York, 2013, vol. 103, DOI: 10.1007/978-1-4614-7138-7, ISBN: 978-1-4614-7137-0 978-1-4614-7138-7. [Online]. Available: http://link.springer.com/10.1007/978-1-4614-7138-7.

[34] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection", *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003. [Online]. Available: http://www.jmlr.org/papers/v3/guyon03a.html.

[35] R. Nilsson, J. M. Peña, J. Björkegren, and J. Tegnér, "Consistent Feature Selection for Pattern Recognition in Polynomial Time", *J. Mach. Learn. Res.*, vol. 8, pp. 589–612, Dec. 2007, ISSN: 1532-4435. [Online]. Available: http://dl.acm.org/citation.cfm?id=1314498.1314519.

[36] M. B. Kursa, W. R. Rudnicki, *et al.*, *Feature selection with the Boruta package*. Journal, 2010. [Online]. Available: https://core.ac.uk/download/files/153/6340269.pdf.

[37] M. A. Hearst, "Support Vector Machines", *IEEE Intelligent Systems*, vol. 13, no. 4, pp. 18–28, Jul. 1998, ISSN: 1541-1672. DOI: 10.1109/5254.708428. [Online]. Available: http://dx.doi.org/10.1109/5254.708428.

[38] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 1st. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1994, ISBN: 978-0-02-352761-6.

[39] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators", *Neural Networks*, vol. 2, no. 5, pp. 359–366, Jan. 1989, ISSN: 0893-6080. DOI: 10.1016/0893-6080(89)90020-8. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0893608089900208.

[40] J. R. Quinlan, "Induction of Decision Trees", *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986, ISSN: 0885-6125. DOI: 10.1023/A:1022643204877. [Online]. Available: http://dx.doi.org/10.1023/A:1022643204877.

[41] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993, ISBN: 978-1-55860-238-0.

[42] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*, en. Taylor & Francis, Jan. 1984, ISBN: 978-0-412-04841-8.

[43] T. G. Dietterich, "Ensemble Methods in Machine Learning", en, in *Multiple Classifier Systems*, ser. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, Jun. 2000, pp. 1–15, ISBN: 978-3-540-67704-8 978-3-540-45014-6. DOI: 10.1007/3-540-45014-9_1. [Online]. Available: https://link.springer.com/chapter/10.1007/3-540-45014-9_1.

[44] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting", *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, p. 1612, 1999.

[45] L. Breiman, "Random Forests", *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, ISSN: 0885-6125. DOI: 10.1023/A:1010933404324. [Online]. Available: https://doi.org/10.1023/A:1010933404324.

[46] T. Ramraj and R. Prabhakar, "Frequent Subgraph Mining Algorithms – A Survey", *Procedia Computer Science*, Graph Algorithms, High Performance Implementations and Its Applications (ICGHIA 2014), vol. 47, pp. 197–204, Jan. 2015, ISSN: 1877-0509. DOI: 10.1016/j.procs.2015.03.198. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050915004664.

[47] X. Yan and J. Han, "gSpan: Graph-based substructure pattern mining", in *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, 2002, pp. 721–724. DOI: 10.1109/ICDM.2002.1184038.

[48] N. Lavrac and S. Dzeroski, *Inductive Logic Programming: Techniques and Applications*. New York, NY, 10001: Routledge, 1993, ISBN: 978-0-13-457870-5.

[49] S. Muggleton and L. De Raedt, "Inductive logic programming: Theory and methods", *The Journal of Logic Programming*, vol. 19, pp. 629–679, 1994.

[50] A. Zheng, *Evaluating Machine Learning Models*, Oct. 2015. [Online]. Available: https://www.oreilly.com/ideas/evaluating-machine-learning-models.

[51] M. A. Hall, "Correlation-based feature selection for machine learning", PhD thesis, The University of Waikato, 1999. [Online]. Available: https://www.lri.fr/~pierres/donn%E9es/save/these/articles/lpr-queue/hall99correlationbased.pdf.

[52] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, Oct. 2011. [Online]. Available: http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.

[53] R. Ihaka and R. Gentleman, "R: A Language for Data Analysis and Graphics", *Journal of Computational and Graphical Statistics*, vol. 5, pp. 299–314, Sep. 1996. DOI: 10.2307/1390807.

[54] J. C. Costello, L. M. Heiser, E. Georgii, M. Gönen, M. P. Menden, N. J. Wang, M. Bansal, M. Ammad-ud-din, P. Hintsanen, S. A. Khan, J.-P. Mpindi, O. Kallioniemi, A. Honkela, T. Aittokallio, K. Wennerberg, J. J. Collins, D. Gallahan, D. Singer, J. Saez-Rodriguez, S. Kaski, J. W. Gray, and G. Stolovitzky, "A community effort to assess and improve drug sensitivity prediction algorithms", *Nature biotechnology*, vol. 32, no. 12, pp. 1202–1212, Dec. 2014, ISSN: 1087-0156. DOI: 10.1038/nbt.2877. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4547623/.

[55] F. Iorio, T. A. Knijnenburg, D. J. Vis, G. R. Bignell, M. P. Menden, M. Schubert, N. Aben, E. Gonçalves, S. Barthorpe, H. Lightfoot, T. Cokelaer, P. Greninger, E. van Dyk, H. Chang, H. de Silva, H. Heyn, X. Deng, R. K. Egan, Q. Liu, T. Mironenko, X. Mitropoulos, L. Richardson, J. Wang, T. Zhang, S. Moran, S. Sayols, M. Soleimani, D. Tamborero, N. Lopez-Bigas, P. Ross-Macdonald, M. Esteller, N. S. Gray, D. A. Haber, M. R. Stratton, C. H. Benes, L. F. A. Wessels, J. Saez-Rodriguez, U. McDermott, and M. J. Garnett, "A Landscape of Pharmacogenomic Interactions in Cancer", *Cell*, vol. 166, no. 3, pp. 740–754, Jul. 2016, ISSN: 0092-8674. DOI: 10.1016/j.cell.2016.06.017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0092867416307462.

[56] W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, S. Ramaswamy, P. A. Futreal, D. A. Haber, M. R. Stratton, C. Benes, U. McDermott, and M. J. Garnett, "Genomics of Drug Sensitivity in Cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells", en, *Nucleic Acids Research*, vol. 41, no. D1, pp. D955–D961, Jan. 2013, ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gks1111. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gks1111.

[57] I. Pilászy and D. Tikk, "Recommending New Movies: Even a Few Ratings Are More Valuable Than Metadata", in *Proceedings of the Third ACM Conference on Recommender Systems*, ser. RecSys '09, New York, NY, USA: ACM, 2009, pp. 93–100, ISBN: 978-1-60558-435-5. DOI: 10.1145/1639714.1639731. [Online]. Available: http://doi.acm.org/10.1145/1639714.1639731.

[58] F. Pereira, P. Norvig, and A. Halevy, “The Unreasonable Effectiveness of Data”, *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8–12, 2009, ISSN: 1541-1672. [Online]. Available: doi.ieeecomputersociety.org/10.1109/MIS.2009.36.

[59] M. Banko and E. Brill, “Scaling to Very Very Large Corpora for Natural Language Disambiguation”, *Microsoft Research*, Jan. 2001. [Online]. Available: https://www.microsoft.com/en-us/research/publication/scaling-to-very-very-large-corpora-for-natural-language-disambiguation/.

[60] A. Maunz, *Data-yeast-ac: NCI Yeast Anticancer Drug Screen: The large scale data used in "Large-Scale Graph Mining using Backbone Refinement Classes"*, original-date: 2008-11-21T15:13:59Z, May 2016. [Online]. Available: https://github.com/amaunz/data-yeast-ac.

[61] S. Min, B. Lee, and S. Yoon, “Deep learning in bioinformatics”, *Briefings in bioinformatics*, vol. 18, no. 5, pp. 851–869, 2017.